PHYSIOLOGICAL CHARACTERIZATION OF PARABELT AUDITORY CORTEX IN THE AWAKE, BEHAVING MARMOSET MONKEY

by

Darik W Gamble

A dissertation submitted to Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy.

Baltimore, Maryland

January 2020

Abstract

The current working model of primate auditory cortex (AC) comprises three successive hierarchically organized stages. The primary 'core' and secondary 'belt' fields are tonotopically organized, but little is known about the topography or physiology of the tertiary 'parabelt' areas. We performed high density single electrode mapping while manually optimizing stimulus properties for well isolated single units, in the awake and behaving marmoset monkey auditory cortex, including parabelt. We observed robust, reliable responses from most units with a relatively simple set of spectrally restricted sounds. The reconstruction of recording site locations exhibited clear evidence that parabelt exhibits a high frequency tonotopic reversal near the anatomically expected border between rostral and caudal parabelt. Receptive fields were consistent with a hierarchical model where, compared to belt auditory cortex, parabelt neurons exhibited longer response latencies, preferred larger bandwidth stimuli, and were more sensitive to temporal modulation.

To further investigate how non-primary auditory cortex represents complex sound stimuli, we characterized neural receptive fields with auditory textures, a rich, diverse class of sounds commonly generated by environmental sources such as rain or wind. Previous work has shown that noise synthesized with statistical structures matched to real-world exemplars are perceived very similarly to their original sound categories, suggesting that statistical summary structure may underlie the perception of auditory texture. We investigated whether the same representation could also explain neural

responses to texture stimuli in higher-level auditory cortex. We developed a novel synthetic texture parametrization and synthesis algorithm that allowed us to synthesize stimuli during an online, closed-feedback stimulus optimization procedure based on genetic optimization. A linear classifier trained on model population responses performed well on sounds synthesized to match real world textures, but more poorly when different statistical structures were removed from the synthesis procedure, a pattern of results that closely matched what has been observed in human psychophysical classification experiments. This suggests that single neurons in higher-level auditory cortex may indeed represent the time-averaged statistical summary of sounds. Our results establish the basic physiology and organization of parabelt areas and confirm their tertiary position in the auditory cortex hierarchy.

Primary Reader and Advisor: Dr. Xiaoqin Wang

Secondary Reader: Dr. Charles E. Connor

Table of Contents

List of Tables

List of Figures

1. Introduction

How do our brains make sense of the acoustic world? Sounds from multiple simultaneous sources arrive together at our ears and mix to stimulate the basilar membrane of the cochlea. The auditory nervous system must then parse this confusion of acoustic signals and segregate information streams of interest from irrelevant background noise, first at the level of basic acoustic properties, and later by higher order spectrotemporal patterns. (Bregman, 1990) The anatomical architecture of the auditory cortex reflects this abstractive process, with connection patterns suggesting a cortical hierarchy where information is serially processed through primary (core), secondary (belt), and tertiary (parabelt) auditory cortical regions. (Jon H. Kaas & Hackett, 2000) There is some physiological evidence for this model: neural responses in core cortical field R remained robust after A1 ablation, but responses in caudal belt regions were diminished, suggesting a dependence on A1 inputs. (J. P. Rauschecker, Tian, Pons, & Mishkin, 1997) Primary auditory fields respond best and most robustly to pure tone stimuli, i.e. the narrowest possible stimulation of the peripheral receptors, whereas neurons in the belt auditory cortex prefer bandpass noise stimuli, an increase in preferred stimulus bandwidth that may reflect the integration of multiple frequencies from primary cortical neurons. (J. P. Rauschecker & Tian, 2004; J. Rauschecker, Tian, & Hauser, 1995; G H Recanzone, 2000; Gregg H Recanzone, Engle, & Juarez-Salinas, 2011; Tian, Reser, Durham, Kustov, & Rauschecker, 2001) In addition to preferences for larger spectral bandwidths, there may also be preferences for temporal modulation:

neurons in belt regions are more selective for linear frequency sweeps. (Tian & Rauschecker, 2004)

In addition to the core/belt/parabelt model of cortical hierarchy, several studies have suggested a bifurcation of information flow in auditory cortex into separate dorsal and ventral streams (J H Kaas & Hackett, 1999; Romanski, Tian, et al., 1999; Romanski, Bates, & Goldman-Rakic, 1999), in analogy to similar models in visual cortex. (Goodale & Milner, 1992; Mishkin, Ungerleider, & Macko, 1983) This interpretation is largely based on anatomical studies that show that caudal belt and parabelt regions predominantly project to cortical regions dedicated to spatial processing in the parietal and dorsal prefrontal cortex, whereas rostral belt and parabelt regions predominantly project to ventral prefrontal regions. There is also some limited physiological evidence for different levels of spatial processing in rostral and caudal regions, but investigation has been primarily limited to the core and belt regions of auditory cortex, as opposed to parabelt auditory cortex. In one study in macaque lateral belt, caudolateral belt field CL had the sharpest spatial tuning, and anterolateral field AL the broadest, with mediolateral ML in the middle. CL also had the lowest selectivity for conspecific vocalizations, and AL had the highest. (Tian et al., 2001) In another study of a larger number of core and belt fields, CL had the sharpest spatial tuning. (Woods, Lopez, Long, Rahman, & Recanzone, 2006) Neurons in macaque caudomedial belt field CM were more spatially sensitive than in A1, (G H Recanzone, Guard, Phan, & Su, 2000) and neurons in marmoset CM/CL were more spatially sensitive than in A1. (Remington & Wang, 2019; Zhou & Wang, 2012) One recent fMRI investigation in macaque

auditory cortex did not find any cortical specialization for spatial processing, (Ortiz-Rios et al., 2017) but another found specializations for auditory motion processing in posterior belt and parabelt regions. (Poirier et al., 2017)

What happens beyond the belt stage? Much less is known about physiological responses in the parabelt auditory cortex compared to earlier auditory areas. Early hypotheses based on anatomical connections and extrapolations from lower levels of cortex suggested that parabelt would preferentially respond to complex, ethologically relevant stimuli such as conspecific vocalizations, or spectrally broad noise. Selectivity for vocalizations has been found in non-primary regions of monkey auditory cortex, but only in regions beyond parabelt auditory cortex, towards the rostral temporal regions. (Perrodin, Kayser, Logothetis, & Petkov, 2011; Petkov et al., 2008; Poremba et al., 2004; Sadagopan, Temiz-Karayol, & Voss, 2015) A recent fMRI study in macaques contrasting passive auditory stimulation with either a random noise background alone, or the same background with a coherent foreground target found selective activation for the target-containing stimuli throughout non-primary auditory cortex—most strongly in rostral parabelt and rostrolateral belt (RLT) but also in caudal parabelt and anterolateral belt areas. An optical intrinsic imaging and electrocorticography (ECoG) study in marmosets investigating tonotopy in auditory cortex suggested there may be an area responsive to high frequencies near the putative border between rostral and caudal parabelt. (Tani, Abe, Hayami, Banno, & Miyakawa, 2018)

In this study, we hypothesized that behavioral engagement in an auditory task would increase neural responses in non-primary auditory cortex to traditional auditory stimuli such as amplitude modulated tones and bandpass noise. We explored the belt and parabelt regions of auditory cortex with single electrode penetrations in marmosets trained to perform an auditory oddball task. This task allowed us to use the same stimuli in and out of a behavioral context, allowing us to both quantify the effects of behavioral engagement on neural response as well characterize neural receptive fields with traditional acoustic stimuli for preferred frequency. By using single electrodes we were able to achieve dense mapping of non-primary auditory cortex and hand-tailor optimized stimuli to each neuron's receptive field, in order to reconstruct tonotopic organization in parabelt auditory cortex.

The results from Chapters 3 and 4 led us to hypothesize that, beyond merely being useful for investigating tonotopy, the selectively for temporally modulated bandpass noise in parabelt might be reflective of a more functional role in the representation of background sounds. Based on this hypothesis we develop, in Chapters 5 and 6, a novel method for characterizing non-primary auditory neurons with synthetic auditory texture stimuli, a class of stimuli particularly well suited for describing environmental background noise.

2. General Methods

2.1 Implantation

Implantation followed previously described methods. (Lu, Liang, & Wang, 2001) All experimental procedures were approved by the Johns Hopkins University Animal Use and Care Committee. Under sterile conditions, with the animal deeply anesthetized with isoflurane (0.5–2.0%), two stainless steel headposts were attached to the skull under sterile conditions and covered with dental acrylic. During implantation, the lateral sulcus was visible through the skull and marked with a pencil for use as a landmark. Recording chambers were built over the lateral sulcus on both sides of the skull, although only the left side was used. After implantation, the animal was allowed to recover for approximately six weeks before returning to food restriction, and then retrained on the same behavioral tasks described above, in a head-fixed condition.

2.2 Physiological experiments

2.2.1 Recording conditions

Experiments were carried out in a double-walled soundproof chamber (Industrial Acoustics, Bronx, NY), with an interior covered by 3-inch acoustic absorption foam (Sonex, Illbruck). The animal was head-fixed in the chair. A micromanipulator (Narishige) mounted drill was used with a 1.0- or 1.1-mm drill bit to drill a craniotomy. The drill angle was fixed at 60 degrees. Any remaining chips of bone were removed with a needle under a surgical microscope. A photograph was taken of the recording

chamber to record the position of the craniotomy. Each day, a tungsten microelectrode (AM Systems, 3-12 MΩ) was loaded into an electrode holder and mounted in a hydraulic microdrive (Trent Wells) in the micromanipulator. The electrode was positioned over a new recording location on the surface of the dura. This position was recorded in a lab notebook. The microdrive was used to push the electrode though the dura and into the cortex. The signal from the electrode was amplified (10,000×) and band-pass filtered (300 Hz – 7kHz) and played over a monitor speaker. The electrode was advanced slowly and in stages until an action potential was detected. Action potentials were sorted online using a template-based spike-sorter (MSD, Alpha Omega Engineering). Because of the selectivity of most neurons, we found that search stimuli were not particularly useful, and thus relied primarily on spontaneous firing to find neurons. The slow, stepwise electrode movement helped ensure we minimized the number of neurons that were ignored due to low spontaneous rates. The depth of each unit was recorded relative to the surface of the dura. The depth at which the first neuron in a penetration was found, relative to the surface, tended to increase, presumably because the dura mater generally thickened due to repeated electrode penetrations. We therefore used the depth of each neuron relative to the first neuron in a penetration as a stable estimate of its cortical depth. Stimuli were synthesized in MATLAB (Mathworks, Natick MA) using custom programs and played from a digital to analog processer (RX6, Tucker Davis Technologies Inc., Alachua, FL), digitally attenuated (PA5, TDT), amplified (Crown D75A), and delivered to speakers via multiplexer (PM2Relay, TDT).

2.2.2 Speaker layout

Speakers (FT2D, Fostex) were arranged in two rings at 0 and 45 degrees elevation, or a single speaker overhead at 90 degrees elevation, as illustrated in Figure 1. These speakers had a flat frequency response between 2-32 kHz. The front speaker at 0 degrees azimuth, 0 degrees elevation was a B&W 601 speaker used because it had a broader response range and was used for neurons with preferred frequencies below 2 kHz. All speakers were positioned 85 cm from the center of the animal's head. The speaker arrangement is shown in Figure 1.

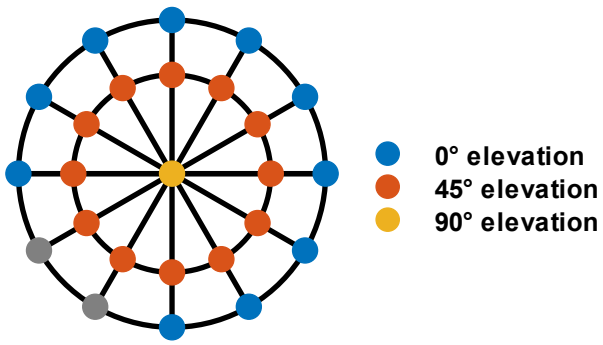


Figure 1 - Recording chamber speaker arrangement

Speakers are shown from an overhead view. The two greyed out speakers in the rear were not present due to the presence of a surgical microscope that prevented their use.

2.2.3 Delineation of cortical field borders

Neurons were assigned to auditory cortical fields using a confluence of receptive field properties and location. Core areas were identified by proximity to the lateral sulcus, short response latencies, and robust responses to pure tone stimuli. The transition to belt was marked by longer response latencies and preference for bandpass noise over pure tones. The transition to parabelt was marked by an abrupt change in tonotopy (See

Chapter 4.) The lateral/inferior border of parabelt was marked by a lack of response to auditory stimulation and responses to visual stimuli.

3. Physiological characterization of parabelt auditory cortex

3.1 Introduction

In this study we used single electrode penetrations to record well-isolated single units in non-primary auditory cortex. This approach allowed us to optimize stimuli, one acoustic dimension at a time, to find the stimulus that maximized each neuron's firing rate. (Wang, Lu, Snider, & Liang, 2005) This approach was useful for several reasons. First, by optimizing non-spectral acoustic features such as intensity, speaker location, and temporal modulation rate, we were able to measure frequency preferences in the most reliable way, which is important for later reconstructing tonotopic organization. Secondly, by constructing stimulus sets that maximally spanned the neuron's dynamic range of firing rates, we could best measure the effect of behavioral engagement on the modulation of neural firing rates. (See Chapter 4.) Finally, because this approach was strongly modeled after earlier work in marmoset primary cortex (e.g. Bendor & Wang, 2008; Lu et al., 2001) it would simplify comparisons between receptive field measurements in these areas.

3.2 Methods

Receptive fields were constructed for each neuron along the following dimensions:

Best bandwidth: The best bandwidth was defined as the bandwidth of the stimulus that evoked the highest firing rate in the bandwidth tuning function. The preferred bandwidth was defined as the firing rate weighted average bandwidth of all stimuli in the bandwidth tuning function.

Best frequency: The best frequency (BF) of a neuron was defined as the weighted average of the tuning function:

$$BF = exp\left[\frac{\sum \bar{r}_i \, log \, f_i}{\sum \bar{r}_i}\right] \quad (1)$$

where $\bar{r}_i$ is the average response to the i-th stimulus, and f_i is the center frequency of the i-th stimulus.

For most neurons, multiple tuning curves were collected during the stimulus optimization procedure. In this case, one tuning curve was chosen to be the 'definitive' tuning curve for the neuron; this was the tuning curve constructed from the stimulus with the narrowest bandwidth, at the lowest intensity, where at least one stimulus in the stimulus set evoked a significant firing rate. Neurons that could not be driven with a stimulus narrower than 1 octave were not included in the tonotopy estimate.

Tuning width: The tuning width (TW) of a neuron was defined as the first absolute moment of the tuning curve about its best frequency:

$$TW = \frac{\sum \bar{r}_i \left| log_2 \frac{f_i}{BF} \right|}{\sum \bar{r}_i} \quad (2)$$

where f_i are the center frequencies of the stimuli used to construct the tuning curve. The width of the stimuli used to construct tuning functions were set to threshold bandwidth.

Minimum response latency: Latency was calculated using a sliding window. (Chase & Young, 2007) First, all repetitions of a single stimulus were collapsed into a single trial. The average spontaneous rate was measured from the pre-stimulus time window, which was either 200 ms for 250 ms duration stimuli, or 500 ms for 500 ms duration stimuli. Next, a 20-ms window was slid over the stimulus time window. The spike count in the time window was compared to the spike count in the pre-stimulus period, and for each timepoint of the sliding window, the probability was calculated that a Poisson process whose rate equaled the spontaneous firing rate would emit a number of spikes equal to or greater than the number in the sliding window. The center of the earliest time window for which this probability fell below a fixed threshold of $p = 0.001$ was considered the response latency for that stimulus; the minimum response latency for a neuron was the minimum response latency over all stimuli. Because of the expected log-normal distribution of response latencies within each region, ANOVA was performed on log-transformed response latencies.

Best modulation frequency: The best modulation frequency (BMF) was calculated by a spike rate weighted average of firing rates in a manner analogous to the best frequency.

Phase synchronization: Phase synchronization was characterized by the vector strength (VS) (Goldberg & Brown, 1969) and then converted to the Rayleigh statistic (2nVS2, where n is the total number of spikes) (Mardia & Jupp, 2000) to assess statistical significance. A value > 13.8 ($p < 0.001$) was considered statistically significant.

Spatial area: We followed Zhou & Wang, 2012 in defining the spatial area (SA) as the fraction of space across which the neuron responded with a firing rate above 50% of the maximum firing rate. If the neuron responded to all speakers equally, then SA = 1; if the neuron responded to only a single speaker, then SA = 0.025.

Best azimuth: While speaker location in our study was a two-dimensional variable (azimuth, elevation), we simplified comparisons of speaker preferences between behaving and passive conditions to only the azimuth. The best azimuth (BA) was calculated as:

$$BA = acos\left[\frac{\sum \bar{r}_i \, cos(az_i)}{\sum \bar{r}_i}\right] \quad (3)$$

where $\bar{r}_i$ is the average response to speaker azimuth az_i.

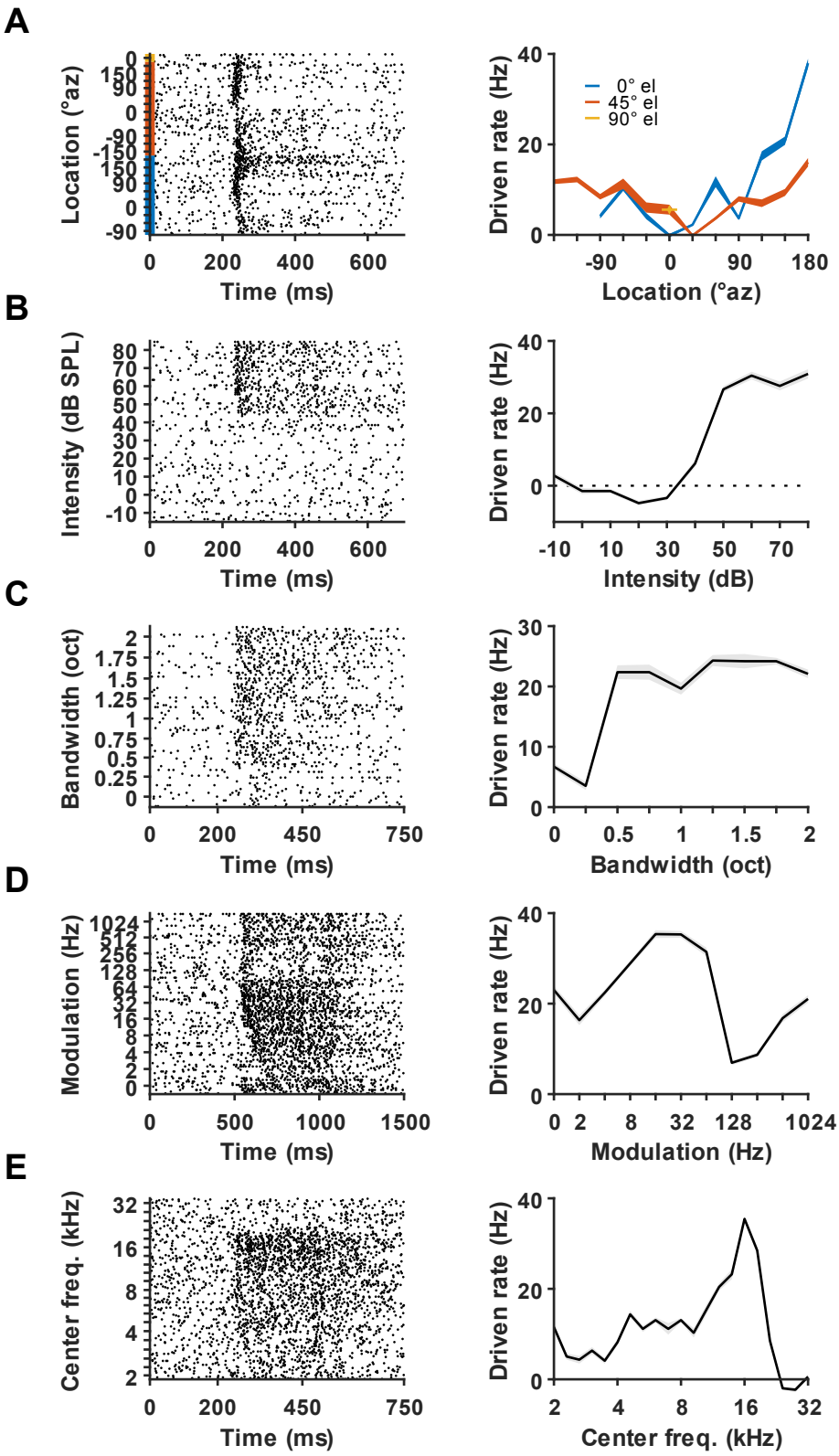
3.3 Results

3.3.1 Stimulus Optimization

Neurons were characterized one at a time by manual stimulus optimization. This process is illustrated in Figure 2 for one example neuron. This neuron's receptive field was iteratively constructed by taking sequential cross-sections through the stimulus space along five dimensions: speaker location, intensity, bandwidth, modulation rate, and finally center frequency. This process allowed us to optimize stimulus parameters enough to be confident that the frequency tuning curve was robust.

Figure 2 - Example stimulus optimization

Stimulus optimization in one example unit. **Left column**: Spike raster plots. Shaded region indicates stimulus duration. Right column: Average firing rates relative to spontaneous rates, mean ± sem. **A**: Spatial receptive field. Spatial tuning was performed with 50 dB broadband noise. This unit responded most strongly to sounds delivered from directly behind the animal's head (180° azimuth, 0° elevation). Successive stimulus sets were all delivered from this location. **B**: Rate-level function. This unit had a threshold of 50 dB SPL. Successive stimulus sets were delivered at this intensity. **C**: Noise bandwidth tuning. Noise was centered at an estimate of best frequency at 16 kHz. The largest bandwidth was limited by the frequency response of the loudspeaker. Successive stimulus sets were delivered at the threshold bandwidth of 0.5 octaves. **D**: Sinusoidal amplitude modulation transfer function. The carrier was the estimate of the optimal stimulus from preceding stimulus sets, i.e. bandpass noise centered at 16 kHz, at 50 dB SPL intensity, delivered from 180° azimuth. Successive stimuli were delivered with this unit's best modulation frequency, 16 Hz. **E**: Frequency tuning. Stimuli were 0.5 octave bandpass noise, at 50 dB SPL intensity, modulated at 16 Hz, whose center frequencies were varied at 5 steps/octave over the frequency response range of the speaker. This unit's best frequency was 16 kHz.


A
Location (°az)
Time (ms)
Driven rate (Hz)
0° el
45° el
90° el
Location (°az)
B
Intensity (dB SPL)
Time (ms)
Driven rate (Hz)
Intensity (dB)
C
Bandwidth (oct)
Time (ms)
Driven rate (Hz)
Bandwidth (oct)
D
Modulation (Hz)
Time (ms)
Driven rate (Hz)
Modulation (Hz)
E
Center freq. (kHz)
Time (ms)
Driven rate (Hz)
Center freq. (kHz)

Throughout belt and parabelt auditory cortex, we found that most neurons could be consistently characterized in this manner and driven with a relatively simple set of bandwidth-restricted stimuli. Figure 3 shows the distributions of threshold and preferred bandwidths—two complementary measures of bandwidth specificity. The threshold bandwidth is the narrowest bandwidth with which we could evoke a reliable response in a neuron. The preferred bandwidth is the bandwidth that evokes the maximum firing rate. In order to characterize the frequency tuning of each neuron we chose stimuli at the threshold bandwidth. Intuitively, if most neurons had very large threshold bandwidths, it would not be very meaningful to discuss preferred frequencies. Conversely, with narrow threshold bandwidths, e.g. one quarter octave, it would be reasonable to quantify frequency tuning. Arbitrarily, we chose one octave as the largest threshold bandwidth for which it was meaningful to discuss a preferred center frequency.

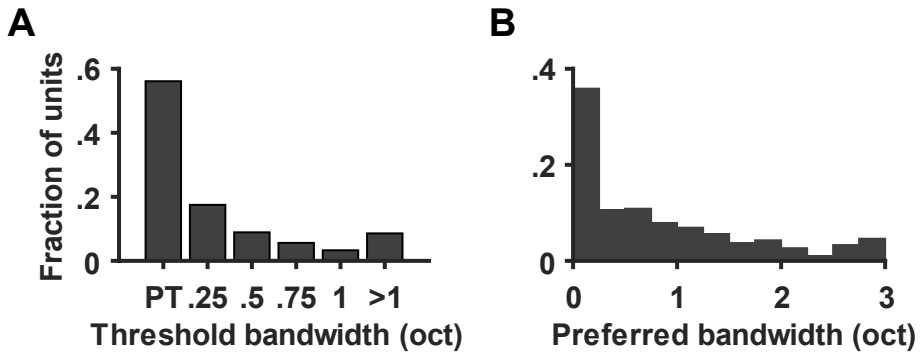


Figure 3 - Distribution of threshold bandwidths.

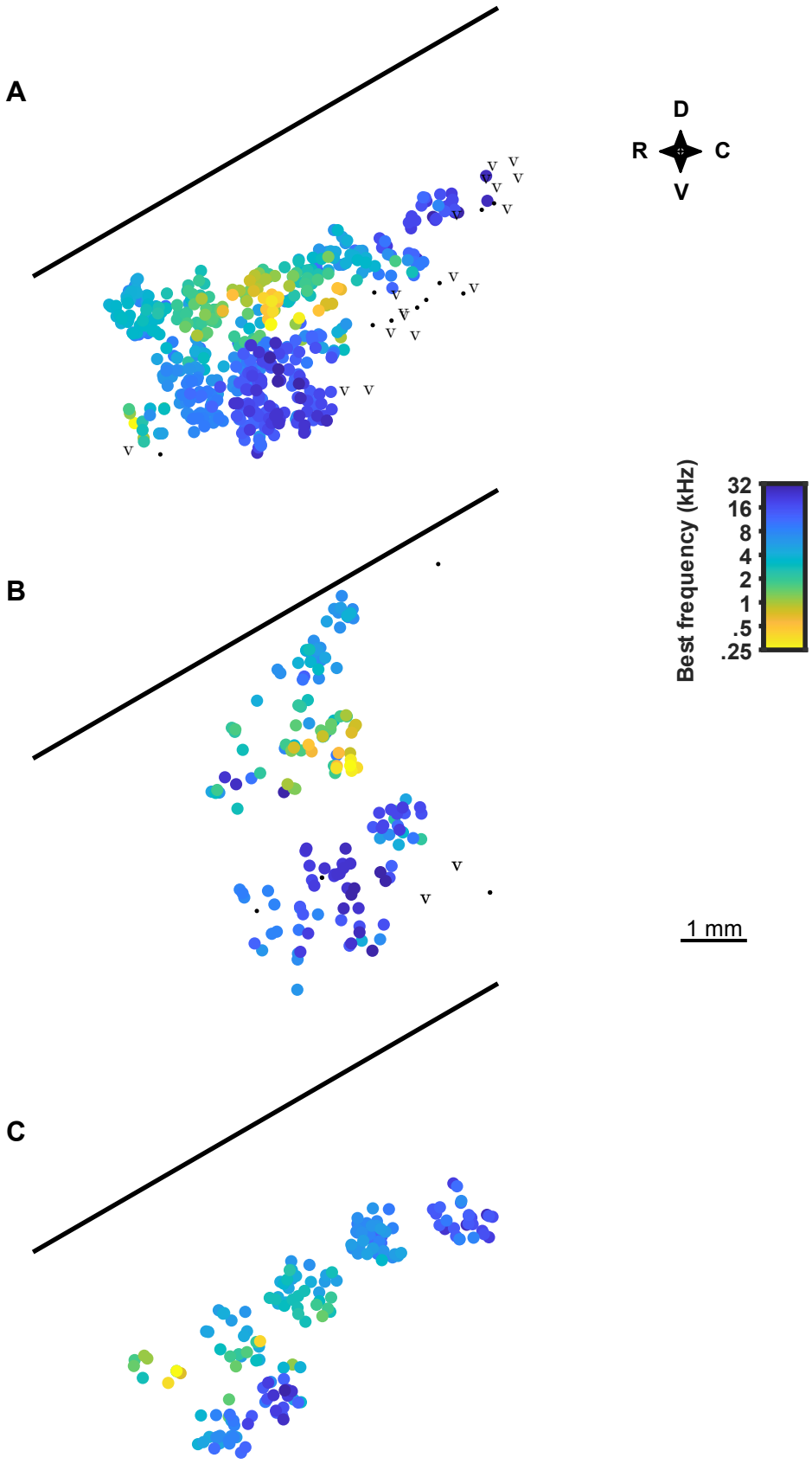
A: Threshold bandwidth, the narrowest stimulus bandwidth that can evoke a significant response. PT: pure tone. Other values indicate bandpass noise bandwidth, in octaves. **B**: Preferred bandwidth. The stimulus bandwidth that evokes the largest response.

3.3.2 Tonotopic Organization

Given a threshold bandwidth, we then swept the center frequency of the stimulus through a wide range of frequencies to construct a tuning curve. From the tuning curve, we obtained both the preferred frequency as well as the tuning width, a measure of frequency selectivity complementary to the threshold bandwidth. In general, it was clear that nearby neurons on the same electrode penetration had similar preferred frequencies. To validate this, we constructed preferred frequencies as a function of the electrode penetration locations, as shown in Figure 4 for three left hemispheres of three animals.

Figure 4 - Reconstructed tonotopic maps

Reconstructed best frequency locations for the left hemispheres of three different animals. **A**: Animal M110z. **B:** M13y. **C:** M118b. The solid black line indicates the lateral sulcus, as marked during the implant surgery. Each point's color indicates the best frequency of a single neuron as shown on the color legend. When more than one unit was recorded from the same electrode penetration, positions were randomly jittered so units from the same track location didn't completely obscure one another. Black 'v' markers indicate electrode penetrations where it was possible to evoke visual responses, and black dots indicate track locations where neither visual nor auditory responses could be identified.


A
B
C
D
R
C
V
Best frequency (kHz)
32
16
8
4
2
1
.5
.25
1 mm

Examination of the reconstructed best frequency maps in Figure 4 reveals several trends. All three maps are consistent with the presence of a high frequency tonotopic reversal lateral to the low frequency reversal known to be present in the core and lateral belt regions. Beyond that we found non-auditory or visual responses consistent with moving outside of the auditory responsive region and into the presumed multimodal areas like the STP and FST. This previously undescribed high frequency tonotopic reversal may reflect the border between caudal and rostral parabelt. If this high frequency tonotopic reversal described previously is in fact part of the parabelt auditory cortex, we would expect to find differences in neural receptive fields there relative to other parts of the auditory cortex. Using the putative borders presented by the best frequency maps, we separated neural populations into tentative regions and compared their stimulus preferences.

3.3.3 Physiological characterization of different cortical regions

3.3.3.1 Neural response latencies

Cortical response latency has been widely considered to be a measurement of hierarchical location; neurons in higher order areas have longer response latencies. (Camalier, D'Angelo, Sterbing-D'Angelo, Mothe, & Hackett, 2012; Schmolesky et al., 1998) Figure 5 shows the distributions of response latencies in the different cortical regions. Qualitatively, response latencies distributions matched the expected pattern,

where neurons in the putative parabelt auditory cortex region had the longest response latencies, and neurons in primary regions had the shortest regions. This supports our earlier hypothesis that the tonotopic patterns observed in the best frequency maps were reflective of borders between different hierarchal levels of the auditory cortex. Because we did not collect data from all three cortical stages in all three subjects, it was not possible to complete a full two factor ANOVA of subject and cortical level. Restricting analysis to only M110Z and M13Y, the subjects for which we had data from both belt and parabelt, ANOVA showed a significant effect for both cortical level ($p < 10^{-5}$), and subject ($p < 10^{-4}$). Collapsing all three subjects, there was still a significant effect of cortical level (Kruskal-Wallis, $p < 10^{-10}$), with all pairwise between-cortical level comparisons significant ($p < 10^{-4}$).

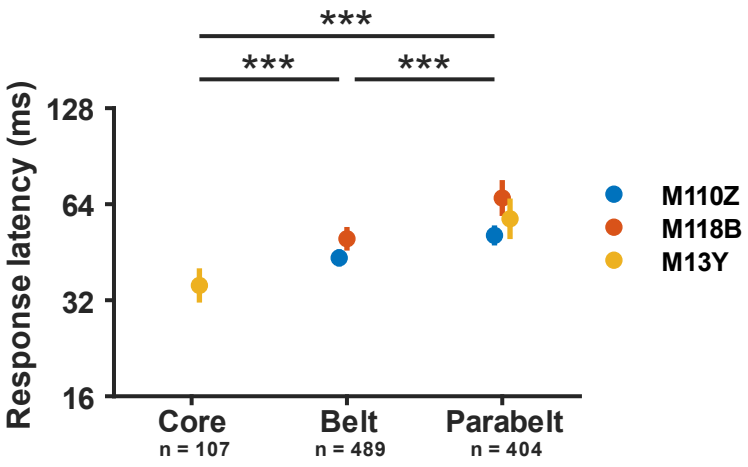


Figure 5 - Distribution of response latencies in different cortical regions

Average response latencies calculated for each animal and cortical level. Numbers below x-axis labels indicate total number of units summed over all subjects.

3.3.3.2 Receptive field bandwidths

Previous reports have found an increase in preferred stimulus bandwidth when moving from core to belt auditory cortex. Do preferred bandwidths continue to increase when moving from belt to parabelt? We tested this in several complementary ways: threshold bandwidth, best bandwidth, and tuning bandwidth. (See methods.) Figure 6 shows the distributions of bandwidths in belt and parabelt regions. For all three measures, parabelt neurons prefer larger bandwidths than belt. There is a significant increase in preferred bandwidth from belt to parabelt, consistent with an overall trend that bandwidth preferences increase while ascending the cortical hierarchy.

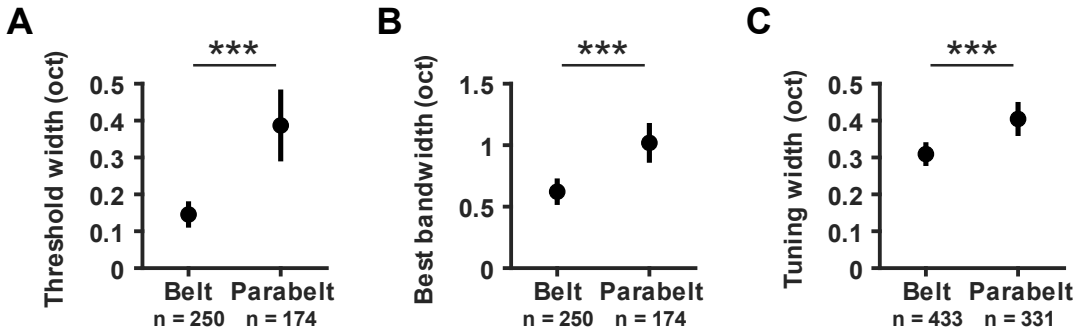


Figure 6 - Bandwidth characterization by cortical level

Comparison of receptive field bandwidths in belt and parabelt regions. **A**: Threshold bandwidth **B**: Best bandwidth. **C**: Tuning curve width. Asterisks indicate ranksum test results ***: $p < 10^{-3}$

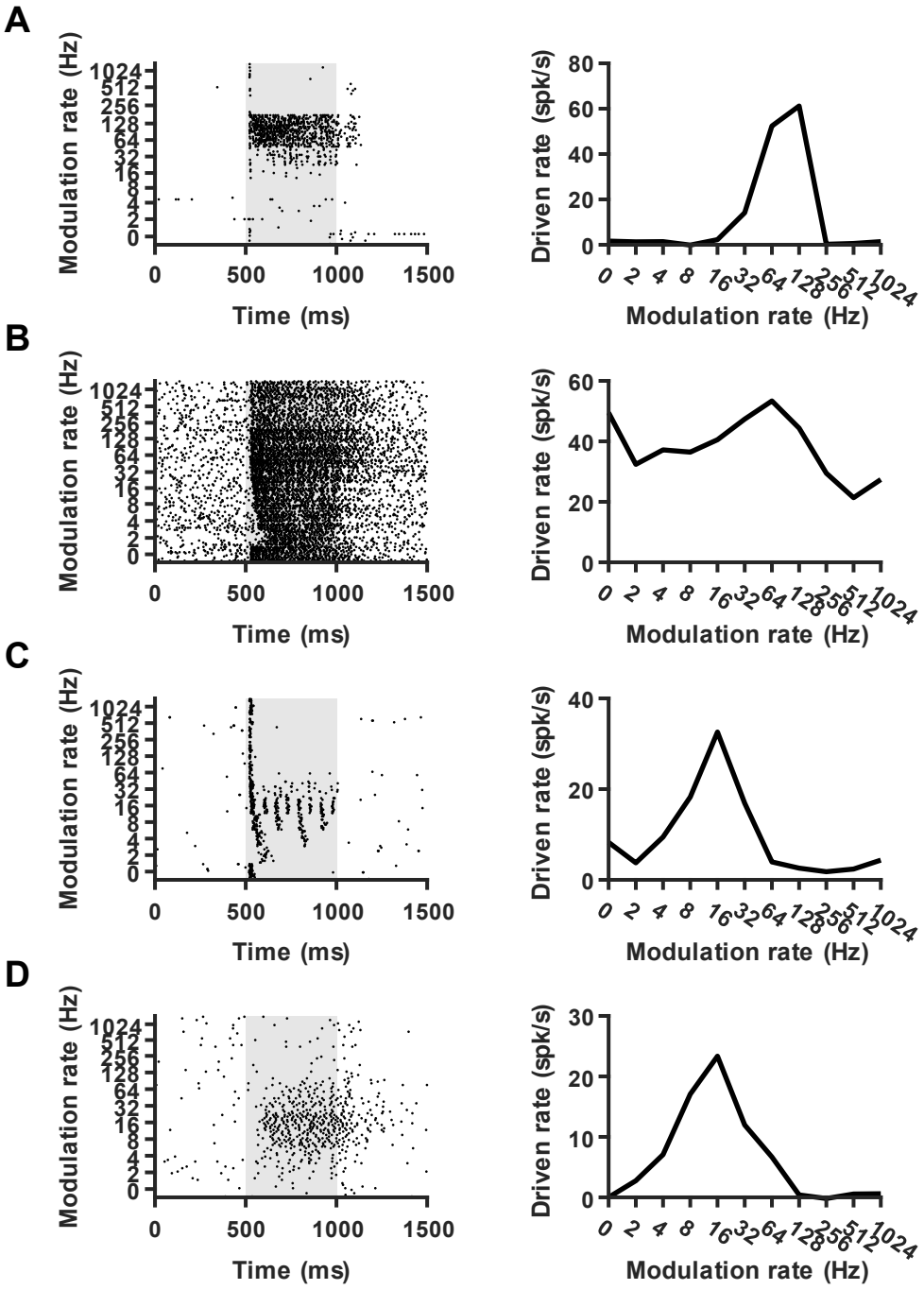
3.3.3.3 Representation of temporal modulation

This increasing bandwidth preference could have multiple interpretations. First, the increase in preferred bandwidth could reflect increasing spectral integration. An alternative hypothesis is that rather than an increase in preferred bandwidth *per se*, there is an increase in preferred stimulus complexity. Another way to increase stimulus complexity is by introducing temporal structure to the noise stimuli envelopes. Given the observed sensitivity to temporal modulation observed in primary auditory cortex, we hypothesized that higher-level auditory cortex may exhibit even more sensitivity to temporal modulation. We tested this first in the simplest possible way, by applying sinusoidal amplitude modulation to the carrier signals. Many neurons responded much more strongly to the modulated signals than to the unmodulated carrier signal alone. Figure 7A shows an example modulation transfer function for a neuron that responded only very weakly to an unmodulated carrier stimulus but responded very robustly to the same stimulus with its preferred modulation rate applied. The unit in Figure 7B, in contrast, responded robustly to the carrier alone and its firing rate was modulated only weakly by amplitude modulation. We characterized this sensitivity to amplitude modulation by the modulation index (See methods). Average modulation index values are shown in Figure 8A. Sensitivity to amplitude modulation is highest in parabelt regions and lowest in core, which suggests that as one ascends the auditory cortical hierarchy, temporal modulation becomes increasingly important for driving neural responses.

While the modulation index is based purely on firing rates, auditory cortical neurons could potentially also use a temporal code. Figure 7C and D show two examples of neurons with similar rate modulation transfer functions, but the unit in Figure 7C exhibits sharp phase locking, whereas the unit in Figure 7D exhibits no significant phase locking. A useful measure for quantifying this is the phase locking index. Figure 8B shows that the maximum rate at which phase locking occurs as well as the fraction of neurons that exhibit significant phase locking at any frequency decreases along the cortical hierarchy. In Figure 8C, we compared the distributions of preferred modulation rates in each cortical level and found no significant differences in modulation rate. Taken together, these results suggest that there is a progression towards the representation of modulation rates by firing rate, and away from temporal representations, as auditory information leaves the core auditory fields.

Figure 7 - Example modulation transfer functions

Example modulation transfer functions for four example neurons demonstrating the diversity of amplitude modulation representation. **Left column**: raster plots: **Right column**: Average firing rates. In each case, the carrier stimuli are bandpass noise stimuli at the units' best center frequency, best level, best bandwidth, and best speaker location. A modulation rate of '0' indicates the unmodulated carrier alone.


A
B
C
D
Modulation rate (Hz)
1024
512
256
128
64
32
16
8
4
2
0
0
500
1000
1500
Time (ms)
Driven rate (spk/s)
80
60
40
20
0
30
10
0 2 4 8 16 32 64 128 256 512 1024
Modulation rate (Hz)

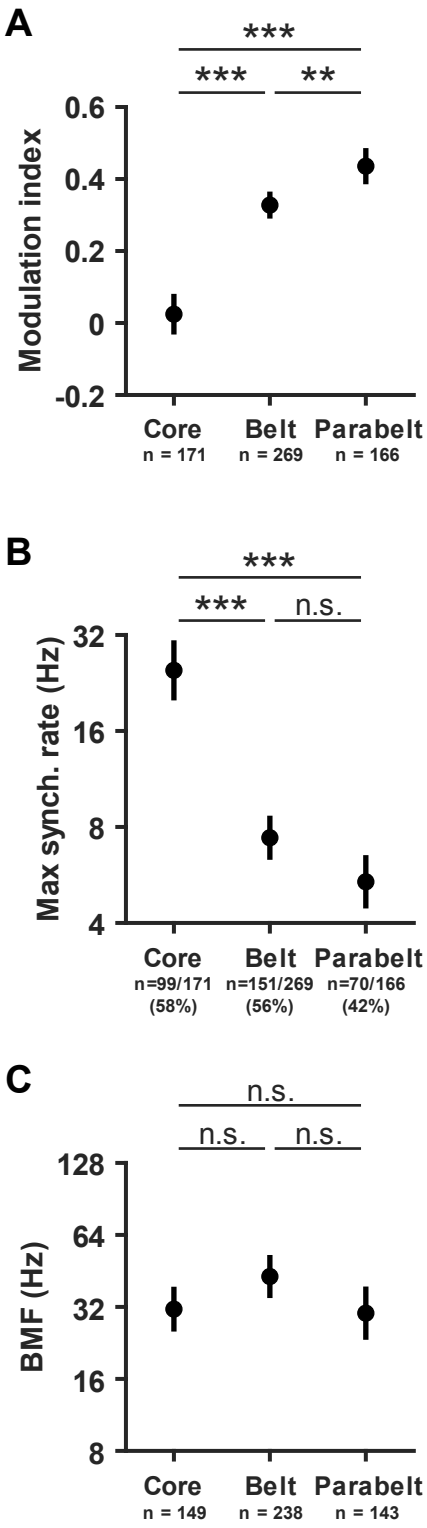


Figure 8 - Summary of amplitude modulation representation.

A: Modulation index values by cortical level. Modulation index values increase along the cortical hierarchy. **B**: Maximum synchronization rate. The maximum synchronization rate is the maximum modulation rate for which a neuron exhibits significant phase locking. The fractions below the x labels indicate which fraction of neurons exhibit significant phase locking at any modulation rate. Both the maximum synchronization rate and the fraction of neurons exhibiting any phase locking decrease across the cortical hierarchy. **C**: Best modulation frequencies (BMF.) Neurons in all three cortical levels have similar preferred modulation rates. In all three subpanels, data for core are from Liang, Lu, & Wang, 2002.

3.3.3.5 Spatial tuning

Sound source location is a necessary component in parsing complex sound scenes. How does the brain represent the acoustic cues required for this processing? Neurons throughout auditory cortex are selective for the spatial location of sound sources. Figure 9 shows four example spatial receptive fields (SRFs) that demonstrate a wide variety of idiosyncrasies from parabelt auditory cortex. The most straightforward way of quantifying the representation of sound location is their spatial selectivity (SA). Highly selective neurons respond to only a highly restricted region of space; more panoramic neurons may respond to sounds from any location. From anatomical studies, it is known that caudal parabelt fields preferentially exchange reciprocal connections with caudal belt regions, and conversely, that rostral parabelt is predominantly reciprocally connected with rostral belt regions. If the dual stream hypothesis were true, one would predict greater spatial selectivity in the 'where' pathway and less selectivity in the 'what' pathway. Figure 10 shows the summary distributions of spatial selectivity in parabelt auditory cortex. Qualitatively speaking, we saw two distinct levels of spatial selectivity. The medial-lateral belt region ML had high spatial selectivity, similar to the selectivity observed in in the caudal belt regions in Zhou & Wang, 2012, consistent with a grouping of ML with CL and CM in a dorsal spatial processing pathway. Surprisingly, we did not observe any difference in spatial selectivity between rostral and caudal parabelt, and furthermore, both fields exhibited low levels of selectivity, approximately equal to that of the anterior-lateral belt region. Thus, it may not be the case that caudal parabelt contributes to spatial processing in the same way as the caudal belt regions do.

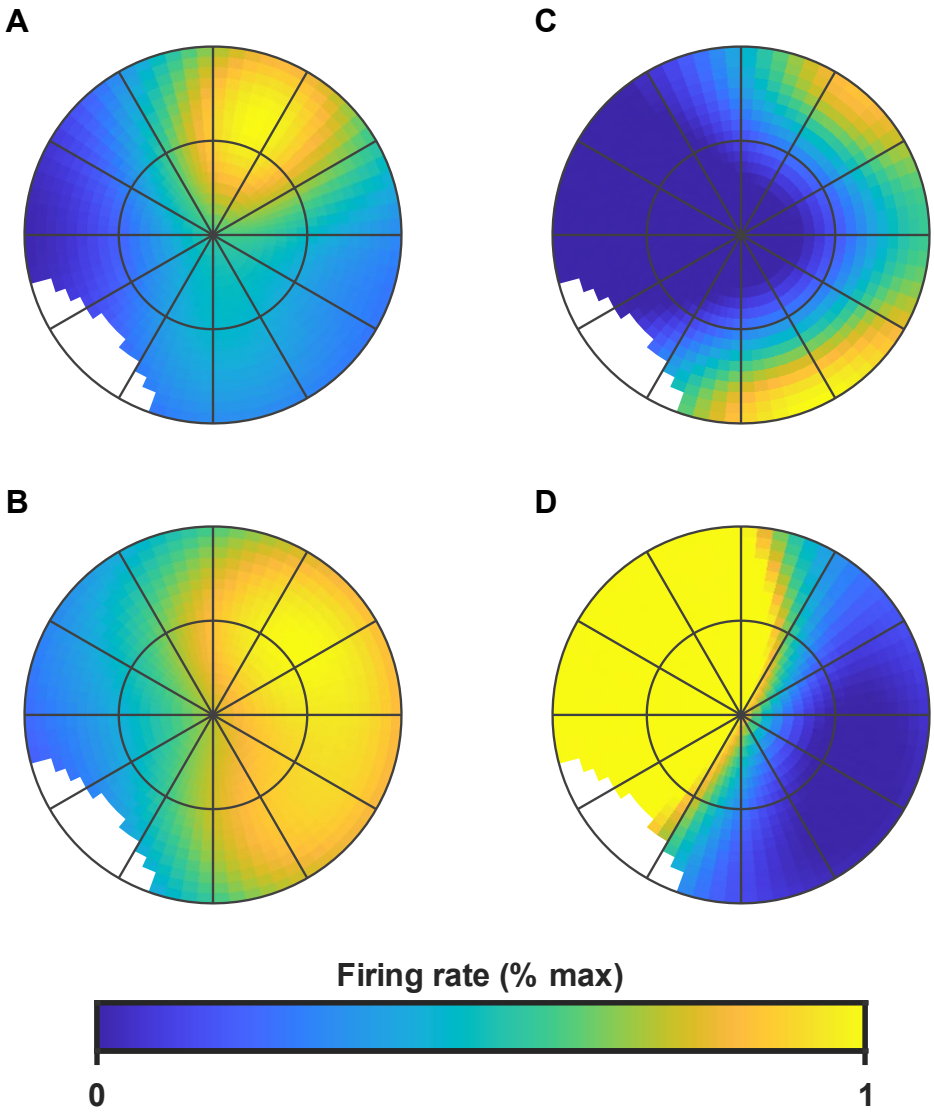


Figure 9 - Example spatial tuning functions

Example SRFs from four different units. SRFs are shown from overhead. (See methods for speaker layout and SRF construction.) **A**: A sharply tuned unit that only responds to sounds from the front-contralateral region. **B**: A unit that responds broadly to most locations in the contralateral hemifield. **C**: A unit that responded to sounds from either the front- or rear-contralateral locations. **D**: A unit that responded broadly to ipsilateral locations.

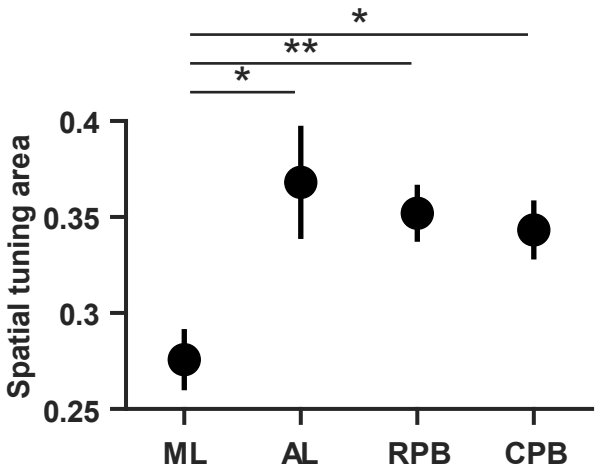


Figure 10 - Summary of spatial tuning areas by cortical field.

Smaller values of spatial tuning area indicate sharper tuning

3.4 Conclusions

Prior hypotheses suggested that parabelt auditory cortex, as the next step after an intermediate processing stage in belt cortex, might be specialized to process ethologically or behaviorally relevant stimuli. While we confirmed a basic hierarchal relationship between belt and parabelt in terms of longer response latencies, larger preferred bandwidth and tuning bandwidths, and increased sensitivity to temporal modulation and decreased phase synchronization, we have found that parabelt neurons can be robustly characterized by a relatively simple set of bandwidth-restricted stimuli, and that parabelt receptive fields represent stimuli in a tuned manner along basic dimensions of acoustic properties such as center frequency, bandwidth, intensity, and temporal modulation. Although we did not systematically test vocalization stimuli, none of the neurons we encountered that could not be driven with this restricted stimulus set showed any robust responses to conspecific vocalizations, consistent with studies that localize vocalization specialization in regions of cortex downstream from parabelt.

By characterizing neurons in non-primary auditory cortex in terms of basic frequency tuning, while keeping other stimulus parameters fixed to optimally drive each unit, we were able to delineate tonotopic organization in the parabelt. This, on one hand, concords with the results in Kajikawa et al., 2015, suggesting that parabelt may be much more responsive to relatively simple acoustic stimuli than previously thought. On one hand, our data are inconsistent with the Kajikawa et al., 2015 observation of a *low-frequency* tonotopic reversal in parabelt, paralleling the adjacent frequency

reversal in belt—we instead observed data consistent with a *high-frequency* reversal; an inversion relative to adjacent belt. Our data are also quite consistent with the intrinsic imaging and ECoG data in marmoset auditory cortex from Tani et al., 2018, which also suggested a high-frequency tonotopic reversal between rostral and caudal parabelt. Currently, we cannot resolve this discrepancy. One possibility is a species-level difference between macaque and marmoset monkeys, but this seems unlikely given how similar otherwise the organization of auditory cortex is in new- and old-world monkeys. The difference could also relate to methodological differences; by using multi-unit activity (MUA) frequency tuning Kajikawa *et al.* were necessarily averaging over larger regions of cortex. Another possibility is that there are more than two fields in parabelt, and each study found different reversals between different pairs of fields. If our high-frequency reversal were to be replicated in other monkey species, it may have interesting implications for the interpretation of human fMRI studies of tonotopic organization, where a long-standing conflict exists between groups that find a tonotopic axis running *parallel* to HG, and others *perpendicular* to HG. (Baumann, Petkov, & Griffiths, 2013; Moerel, De Martino, & Formisano, 2014) This discrepancy may be at least partially due to inter-subject anatomical variability as well as any number of differences in experimental design, but the presence of a high frequency region adjacent to a low frequency could be misidentified as a tonotopic field. Whole brain analysis in a phase-encoded fMRI experiment identified at least six tonotopic maps in human cortex, including four that could not be explained by known core/belt tonotopy. (Striem-Amit, Hertz, & Amedi, 2011) Identifying and characterizing the tonotopic

organization of non-human primate parabelt cortex will play a major role in the interpretation of future human and monkey imaging studies.

In any case, these results, taken together with other work identifying vocalization specializations in regions beyond parabelt, suggest a more intermediate role for parabelt in stimulus representation than might otherwise have been expected. But why is bandpass noise such an effective stimulus in non-primary auditory cortex? Initial hypotheses (J. Rauschecker et al., 1995) drew an analogue to increases in spatial receptive field size in the ascending visual cortex, allowing higher order visual cortex to gradually represent entire objects rather than its constituent parts. By analogy, larger spectral receptive fields in auditory cortex would allow non-primary auditory cortex to represent "spectrally large" sounds. But an immediate criticism of this hypothesis is that bandpass noise stimuli are largely unnatural, especially compared to the harmonically rich acoustic structures present in ethologically relevant sounds like vocalizations. While bandpass noise responses could be purely epiphenomenal, an alternative interpretation is that rather than specialization for behaviorally relevant stimuli, parabelt auditory cortex is particularly suited for representing irrelevant background noise, which is often broadband, non-harmonic, and temporally modulated. This representation of irrelevant background noise could then be used to perform subtractive target isolation.

Contrary to predictions based on the auditory dual steam hypothesis, we did not detect any differences in spatial selectivity between rostral and caudal fields, although we did

see significantly higher spatial selectivity in the more caudal belt field ML, similar to what has been previously described in CM/CL. (G H Recanzone et al., 2000; Remington & Wang, 2019; Tian et al., 2001; Woods et al., 2006; Zhou & Wang, 2012) This does not necessarily contradict tracer studies which were largely based on injections into the caudal belt regions, rather than parabelt, (Romanski, Tian, et al., 1999) although injections in dorsal and ventral prefrontal cortex selectively labeled neurons in caudal and rostral parabelt (Romanski, Bates, et al., 1999), and caudal belt regions preferentially exchange connections with caudal belt, and likewise rostral belt with rostral parabelt regions. (de la Mothe, Blumell, Kajikawa, & Hackett, 2012; Hackett, Stepniewska, & Kaas, 1998) Again, if parabelt was performing a form of background subtraction, then caudal and rostral parabelt could be performing identical roles in the two streams. This would allow caudal belt regions to send target spatial locations to the spatial processing targets in dorsal parietal and pre-frontal cortical regions.

While neurons in belt and parabelt were sensitive to temporal modulation in a similar manner to neurons in core regions, their sensitivity to modulation increased, and synchronization became less common. This parallels the trends observed within core regions, where synchronization decreased from A1 to R to RT. (Bendor & Wang, 2008) Whereas Bendor & Wang 2008 hypothesized two independent dimensions in auditory cortex, with temporal integration windows increasing along a rostrocaudal axis, and spectral integration windows widening along a mediolateral axis, our results here suggest that integration windows broaden along both axes. Contrary to a recent fMRI

investigation in macaque primary auditory cortex (Baumann et al., 2015), we did not observe any topographic representation of modulation rate in auditory cortex.

4. Effect of behavioral engagement on parabelt neural activity

4.1 Introduction

Although the previous chapter addressed neural receptive fields in non-primary auditory cortex as though they were fixed functions of stimulus acoustic properties, responses in higher-level areas might depend not only on stimulus selectivity but also behavioral engagement or attentional selection. Neurons in ferret A1 exhibited shifts in frequency preferences related to target frequencies, (J. B Fritz, Elhilali, & Shamma, 2005; J. Fritz, Shamma, Elhilali, & Klein, 2003; Jonathan B. Fritz, Elhilali, & Shamma, 2007) as well as shifts in modulation preferences. (Yin, Fritz, & Shamma, 2014) Beyond A1, neurons in ferret prefrontal cortex exhibited behavioral gating, (Jonathan B Fritz, David, Radtke-Schuller, Yin, & Shamma, 2010) and neurons in the non-primary auditory cortex area behaved in an intermediate way between A1 and PFC neurons. (Atiani et al., 2014) Spatial selectivity in cat A1 neurons sharpened during a spatial task, relative to a non-spatial task or passive control situation. (Lee & Middlebrooks, 2010) Some neurons in macaque auditory cortex responded to reward expectations. (Brosch, Selezneva, & Scheich, 2005, 2011)

We addressed this issue by training marmosets to perform an auditory odd-ball detection task in a head-fixed condition, so that we could measure neural responses in and out of a behaving task and compare responses between conditions.

4.2 Methods

4.2.1 Behavioral training

Behavioral training followed previously published methods. (Osmanski & Wang, 2011; Remington, Osmanski, & Wang, 2012) First, subjects were weighed every day for approximately 10 days to estimate their free-feeding weight. Next, we began restricting access to food, aiming at a target weight of 90% of the free-feeding weight. This target weight was approached slowly over the course of several days to ensure we did not over-restrict food access. Animals were rewarded with small marshmallows for transferring from their home cage to carrier cages, after being handled with leather gauntlets, and further for sitting in the primate chair. Once sitting in the chair, marshmallow rewards were replaced with the primary reward diet, which was a combination of watered rice cereal, Similac powder, and strawberry Nesquik flavoring. This reward was delivered by a syringe pump (NE-500, New Era Pump Systems Inc., Farmingdale, NY) through flexible plastic tubing into a hard plastic feeding tube. Initially, the goal was to habituate animals to feeding from the tube, as well as to pair the sound of the syringe pump with food delivery.

Training began with the experimenter inside the chamber, in front of the animal. Soon after, the feeding tube was attached to the neck plate of the primate chair, and the experimenter left the recording chamber. Reward sizes were initially large and frequent, and over the course of 3-5 days became small and infrequent. Generally, animals would consume approximately 10-20 mL of reward food, so training sessions

necessarily grew in duration over the course of training. Early in training, animals would often not notice, or ignore food delivery, but after habituating to the chair restraint they quickly consumed any ejected food. This habituation was often accompanied by the emergence of an obvious orientation response to the sound of the food pump. Around this time, or slightly earlier, we began preceding the reward delivery with the presentation of a loud (75 dB SPL, roved +/- 5 dB) temporally orthogonal ripple combination (TORC) stimulus (see below). If the animal licked at the feeding tube during the stimulus (in anticipation of the reward) the sound was silenced, and delivery commenced immediately; otherwise, the sound played to its full duration, and then the reward was given. To gradually increase the possibility of anticipatory responses, either the duration of the sound was increased, occasionally up to 30 seconds, or the number of target sounds within the trial was increased, up to 10 repetitions. Eventually, the non-contingent reward delivery was eliminated, and the animal was rewarded only for responding during the target sound delivery.

At this point, 30% of the trials were replaced with sham silent targets, to estimate the animal's false alarm rate. Once the animal performed with d' > 3 for three consecutive days, responding to three consecutive 250-ms long target stimuli with 750 ms SOA, with 5-15 seconds between trials, the animal was moved to the auditory discrimination stage of training. There, the target stimuli remained the same TORC stimuli, but were presented embedded in random sequences of non-target narrowband noise stimuli with random center frequencies between 2-32 kHz. Initially, there were 3-5 background stimuli preceding each target stimulus, and background stimuli were played quietly,

approximately 20 dB SPL. Gradually over the course of 2-3 weeks, the intensities and number of background stimuli were increased until each trial consisted of 1 target stimuli in a sequence of 12-18 250-ms-long background stimuli with 750 ms SOA, played at the same average intensity as the target stimuli. As before, 30% of the trials were shams; in this case, instead of a silent target, the original TORC stimulus was replaced by a background stimulus. In either case, a response window was set at the onset of the target and terminated 1500 ms after the onset of the target. A response in this window was classified as either a hit or false alarm; responses outside of the window were classified as errors. False alarms invoked an additional 3 second penalty in addition to the standard inter-trial interval; errors early in a trial invoked not only that 3 second timeout, but also an additional timeout equal to the amount of time remaining in the trial; this prevented high error rate sessions from being shorter in total duration than low error rate sessions. Animals were considered to have completed this phase of training after three consecutive days with a d' > 3.

In the final stage of training, the task remained the same but some arbitrary stimulus parameters were varied, to ensure the animal could maintain the same level of performance with different background stimulus sets varying in speaker location, intensity, bandwidth, and modulation rate.

4.2.2 Passive and behaving conditions

Physiological experiments alternated between passive and behaving conditions. Throughout all recording sessions, pupillometry data was recorded with an infra-red

camera (ETL-200, ISCAN Inc., Woburn MA). Pupil position and diameter data were aligned to stimulus presentations by a synchronization pulse between TDT and ISCAN hardware. A single LED, mounted in the chamber at eye level, was turned on during behavior sessions to indicate to the animal when rewards were available.

The behavioral task was an oddball detection task. Background stimuli were chosen from a set of stimuli used in the passive condition that spanned the unit's receptive field in one dimension. (For example, a set of bandpass noise stimuli of varying center frequencies.)

Targets were chosen to be relatively infrequent TORC stimuli, (Klein, Depireux, Simon, & Shamma, 2000) which was useful for three reasons. First, TORCs are perceptually distinct from the stimuli traditionally used for the physiological characterization of auditory receptive fields. This makes the TORC detection task much easier to generalize across different sets of traditional background stimuli. Second, because the traditional use of TORCs for characterizing STRFs through reverse correlation relies on strong phase-locking to the amplitude modulation, they were not particularly useful in this regard in non-primary auditory cortex, where phase-locking is much weaker than in primary areas. TORCs could therefore be used exclusively as targets. Third, because marmosets are somewhat limited in the number of behavioral trials they can perform, by constructing our trials with a high number of background stimuli, and a small number of oddball targets, we maximized the number of presentations of neurophysiologically relevant stimuli. This design was instrumental in

allowing us to test the effect of behavior on not just frequency tuning, but also temporal modulation and spatial tuning, with dense sampling of the receptive field focused on the stimulus ranges that evoked neural responses.

Eye tracking: To determine if the animal's eyes were open or not during stimulus presentation, the pupil diameter data was averaged over the pre-stimulus and stimulus period and compared to a fixed threshold set by manual inspection and confirmed by comparison with recorded raw camera data. Although animals' eyes were predominantly open during behavior sessions, they did occasionally close mid trial; these stimulus presentations were discarded. Further analysis was then performed on stimulus presentations classified into three levels of arousal: 1) passive/non-behaving, eyes closed; 2) passive/non-behaving, eyes open; and 3) behaving, eyes open. We kept any stimulus with at least three repetitions each in at least two categories, and any neuron with at least one stimulus fitting that criteria.

To determine the effect of arousal state on neural stimulus preferences we calculated response weighted stimulus averages as described above for BF, BMF, and BA for each unit and category separately. Because the animal's eye state was not under experimenter control, and there was not always time to present stimulus sets in both the passive and behaving conditions, data for some units were not always available under all conditions. When applying repeated measures ANOVA, (using the Matlab functions *fitrm* and *ranova*) we considered only units and stimuli that were presented in all three conditions.

ANOVA analysis of z-scored neural responses:

We defined r_{ijk} as the average firing rate of neuron i, to stimulus j, over all repetitions in arousal condition k. The mean and variance of the response of neuron i to stimulus j over arousal conditions are then

$$\mu_{ij} = E_k[r_{ijk}] \tag{4}$$

and

$$\sigma_{ij}^2 = Var_k[r_{ijk}] \tag{5}$$

where $E_t[x_t] = \frac{1}{n_t}\sum_t x_t$ is the average over dimension t and $Var_t[x_t]$ is similarly defined.

The z-scored response is therefore

$$z_{ijk} = \frac{r_{ijk} - \mu_{ij}}{\sigma_{ij}} \tag{6}$$

and we obtain the average z-scored response of neuron i in condition k by averaging over all stimuli j for that neuron:

$$z_{ik} = E_j[z_{ijk}] \tag{7}$$

The values of z_{ik} along with which animal and cortical level neuron i was assigned to were input to the MATLAB n-way anova function *anovan*. All three factors (animal, cortical level, and arousal state) were specified as random effects. The effect of behavior on average spontaneous firing rates were calculated in a similar manner, defining s_{ijk} as the average spontaneous firing rate for neuron i in stimulus set j in arousal condition k, and calculating z-scores in the same manner as described above.

4.3 Results

4.3.1 Behavioral performance

We found that after the initial retraining period after implantation, animals performed the oddball detection task easily even when head-fixed, and had few difficulties switching between active behavior and passive listening sessions, even when using novel stimulus sets for background stimuli. Collapsing all behavior sessions and animals, the average performance was a hit rate of 96% and a false alarm rate of 2%. Although animals did occasionally respond during passive conditions, they could clearly distinguish between behaving and passive conditions, since they responded, on average, to 7% of all sound presentations in the behaving conditions, and only 0.5% of stimuli in the non-behaving condition.

4.3.2 Effect of behavior on neural firing rates

In general, behavior engagement and arousal increased the firing rates of neurons throughout non-primary auditory cortex without changing their stimulus preferences. Figure 11 shows example frequency tuning curves in the three different arousal conditions. These example neurons suggest two things: neural stimulus preferences remain consistent between arousal states, but overall gain increases with increasing arousal levels. To first confirm that stimulus preferences remained fixed, we compared best frequencies, best modulation rates, and best speaker azimuths between arousal conditions. Repeated measures ANOVA confirmed that there was no significant effect

of arousal condition on stimulus preferences in terms of BF ($p = 0.87$), BMF ($p = 0.29$), or BA ($p = 0.27$).

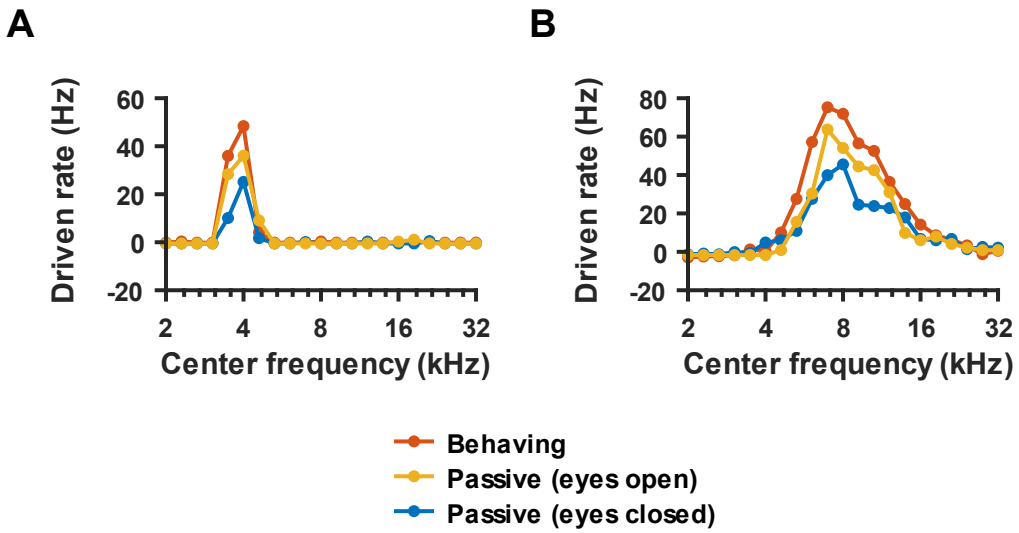


Figure 11 - Effect of behavior on frequency tuning curves.

Example tuning curves for two different example units, one (**A**) narrowly tuned, and one (**B**) broadly tuned. In both examples, the highest firing rates occur in the behaving condition, and the lowest in the passive, eyes closed condition.

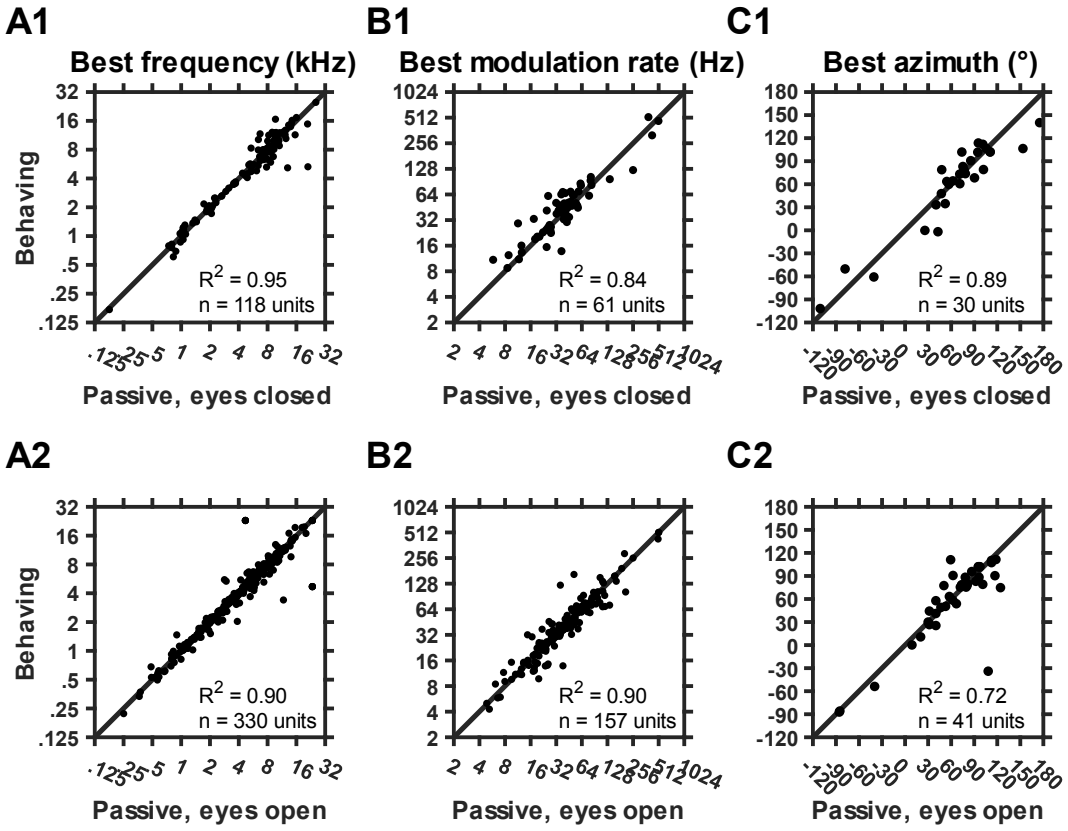


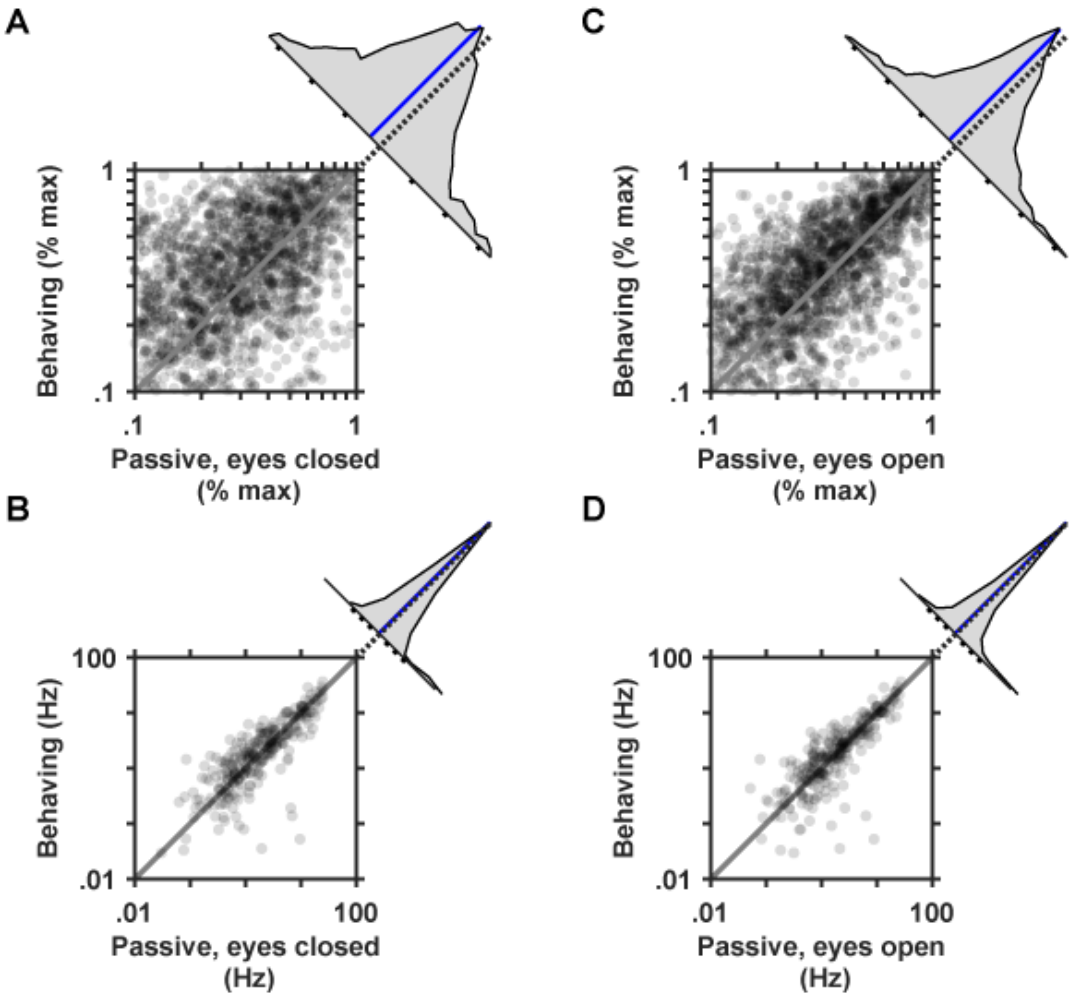
Figure 12 - Receptive field stability between behaving and passive conditions

Top row: Comparisons between the behaving, and passive, eyes closed conditions. **Bottom row**: Same but between the behaving and passive, eyes open condition. The corresponding comparisons between the two passive conditions are similar but not shown. **A1-2:** Best frequency. **B1-2**: Best modulation frequency. **C1-2**: Best azimuth.

To determine the effect of arousal at a population level, we collapsed together all stimuli that evoked a reliable response in at least one condition. Results are shown in Figure 13. Matching the trends observed in the examples in Figure 11, average firing rates in the behaving condition were 23% higher than in the passive, eyes closed state. Consistent with its interpretation as an intermediate level of arousal, firing rates in the eyes open state were 15% higher than in the eyes closed state.

Figure 13 - Stimulus level effect of arousal on firing rates

A: Comparison of stimulus-evoked firing rates between the behaving and passive, eyes closed condition. Each point represents the average firing rate of a single neuron to a single stimulus in two different conditions. Points above the identity line are stimuli whose firing rate was higher in the behaving condition, relative to the firing rate in the passive, eyes closed condition. Markers are partially transparent, so darker regions indicate a higher density of points. Stimuli were included if the neuron responded > 10% max in at least one of the three arousal conditions. Due to differences in sampling density and neural responsivity, different neurons contribute different numbers of points to this plot. The inset histogram shows the distributions of gains (gain = y/x), and the blue line shows the median value of this distribution. The median of this distribution is 1.23, (i.e. an 23% increase in firing rate) which is significantly higher than 1 (ranksum; $p < 10^{-5}$). **B**: Same conventions as in A, but for the behaving and passive eyes open condition. Here the median gain is 1.15 ($p < 10^{-5}$). **C**: Same convention as in A, but for spontaneous firing rates. Each point represents the average spontaneous rate of a neuron collapsed over all sessions for that arousal condition. The gain is not significantly different from 1. (median = 1.06, $p = 0.32$). **D**: Same conventions as C, for the behaving and passive, eyes open condition. Median gain = 1.05, $p = 0.15$.


A
Behaving (% max)
1
.1
.1
1
Passive, eyes closed
(% max)
C
Behaving (% max)
1
.1
.1
1
Passive, eyes open
(% max)
B
Behaving (Hz)
100
.01
.01
100
Passive, eyes closed
(Hz)
D
Behaving (Hz)
100
.01
.01
100
Passive, eyes open
(Hz)

To quantify these effects at a group level, we removed the main effects of neuron and stimulus and considered only the effects of cortical level (belt vs. parabelt), subject, and arousal condition. Mean effects are shown in Figure 14. Three-way ANOVA revealed a significant effect only for arousal condition ($p < 10^{-5}$), but not cortical level ($p = 0.11$) or subject ($p = 0.90$). There was no significant interaction between cortical level and subject ($p = 0.07$). The level of arousal, as indexed by eye opening and behavioral state, seems to be correlated in general with an overall effect of neural gain that acts similarly in both belt and parabelt regions.

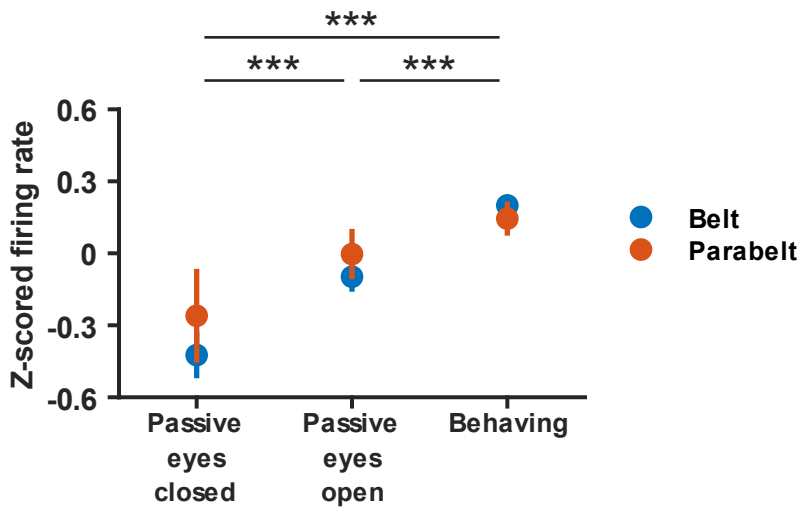


Figure 14 - Mean effect of arousal condition on stimulus evoked firing rates

Values indicate mean z-scored firing rates after removing the main effect of stimulus, averaged over each neuron. Asterisks indicate significance of the main effect of arousal condition.

4.4 Conclusions

Although neurons in parabelt increased their neural gain when in an actively behaving condition, stimulus preferences remained largely fixed, and the increase in neural gain was not any more pronounced in parabelt than in belt, suggesting that this gain may be a general effect of distributed arousal-related neuromodulatory systems, rather than a behavior-specific circuit-level effect. (Harris & Thiele, 2011a; McGinley et al., 2015; Zagha & Mccormick, 2014) The effect we observed was similar to what has been reported in numerous studies in sensory cortex, such as increased firing rates under locomotion in rodent visual cortex (Niell & Stryker, 2010; Polack, Friedman, & Golshani, 2013) or in somatosensory cortex under active whisking (Crochet & Petersen, 2006; O'Connor, Peron, Huber, & Svoboda, 2010)

This supports our approach in the previous chapter of calculating stimulus preferences as the center of mass of stimulus property tuning functions, which would not be affected by a constant gain factor. However, our behavioral design cannot separate contributions of attentional selection and overall effects of behavioral arousal; future studies should develop a multiple auditory stream selective attention task analogous to studies on visual selective attention. (Moran & Desimone, 1985)

5. Low Dimensional Representation of Auditory Textures

5.1 Introduction

'Visual texture' refers to dense, repetitive patterns of small elements we can learn to recognize and distinguish based on their composite appearance. For example, marble is perceptually distinct from wood grain, and we can distinguish these materials categorically. Likewise, tactile textures are dense patterns of small surface features that form textures like silk and sandpaper. By analogy, 'auditory texture' refers to dense, repetitive patterns of short-duration sounds generated by environmental sources; consider, for example, the difference between the sound of rainfall and a waterfall. These sorts of environmental sounds commonly form the background to the auditory scenes we move through every day. They can serve as critical cues about the location, nature, and state of the surrounding world. Often, they mask relevant sounds, such as speech, and must be perceptually subtracted.

Because auditory textures arise from the combination of many independent sources (for example, individual raindrops), they can be more efficiently described by their probability distributions, rather than in terms each individual source. A probabilistic representation of auditory texture was developed by McDermott and colleagues (McDermott, Schemitsch, & Simoncelli, 2013; McDermott & Simoncelli, 2011), building on previous work on descriptions of visual texture. They developed a

technique of iterative noise filtering to synthesize a novel sound token that exhibited the same statistical structure as a specified set of target statistics measured from a real-world exemplar. Sounds synthesized in this way were perceptually similar to the original exemplar sounds, suggesting that statistical structure may indeed underlie the neural representation of auditory texture. Importantly, the statistical summaries were chosen to be relatively simple to calculate, meaning that they could be plausibly implemented by a biological neural network.

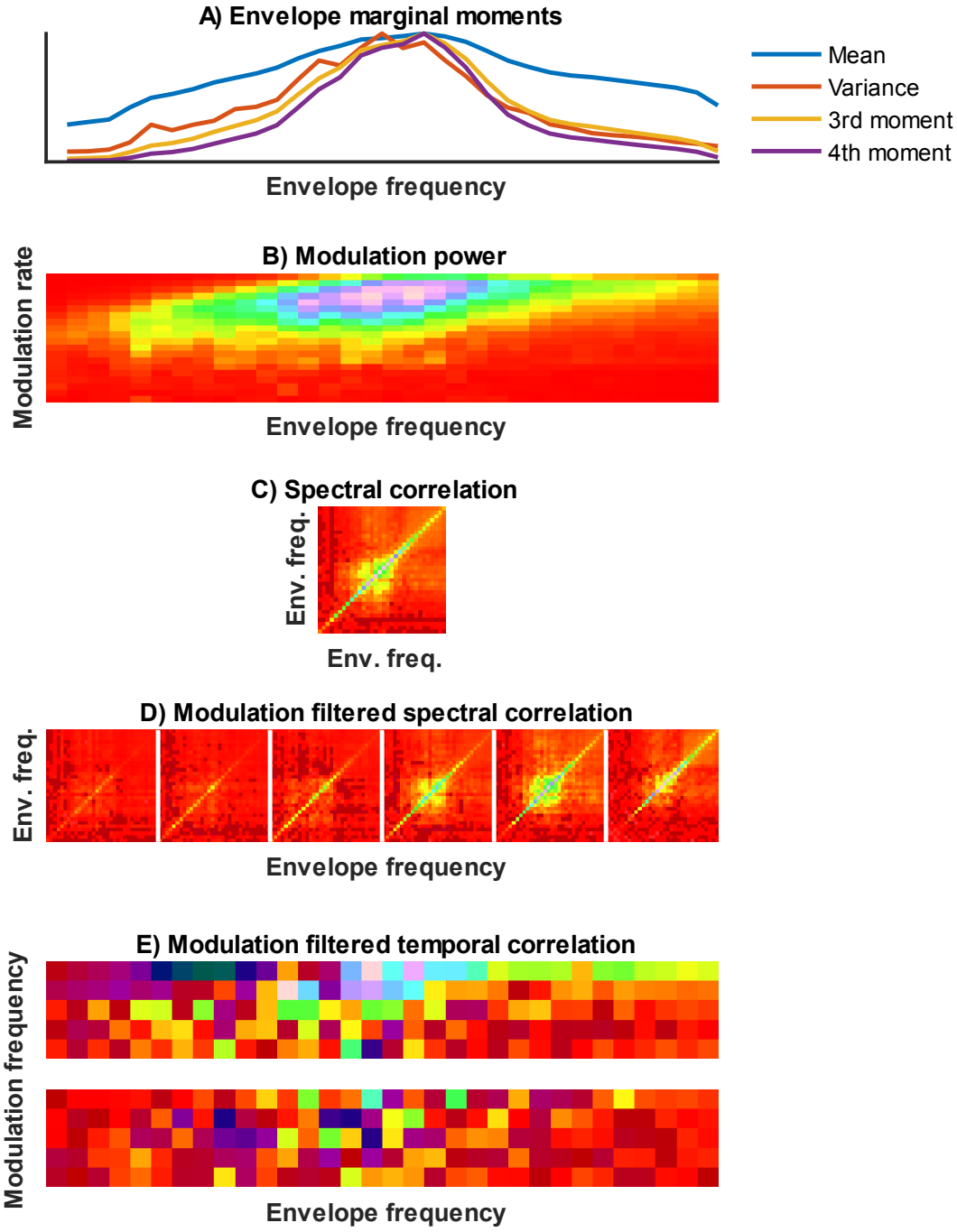
We are interested in using auditory textures to study auditory cortex for several reasons. First, most neurophysiological investigations into the cortical representation of sound rely on relatively simple and unnatural sounds such as pure tones, harmonic tone complexes, click trains, and bandpass noise. Studies utilizing more complex, ethologically relevant stimuli almost exclusively rely on vocalizations—certainly, one of the most important stimulus classes, but far from spanning the entire range of natural sounds. Stimuli of intermediate complexity tend to rely on linear systems theory, such as in ripple sounds and ripple combinations, (Depireux, Simon, Klein, & Shamma, 2001) or random spectral stimuli. (Barbour & Wang, 2003; Yu & Young, 2000). Auditory textures present a novel approach to studying the representation of complex, naturalistic stimuli. They appear particularly suited for studying integrative processing in non-primary auditory cortex, due to their broad frequency spectra and complex temporal structures. Secondly, in addition to their pragmatic utility as an experimental stimulus, by characterizing neural responses to auditory textures we can also address

the basic scientific question of how auditory textures in general, and their statistical structures, are represented in auditory cortex.

There are several major roadblocks to using auditory textures in a neurophysiological experiment. The first difficulty relates to stimulus dimensionality. There is a huge number of textures that could be synthesized, and only a limited amount of experimental time available. By reducing this stimulus space to a lower dimensionality, we hypothesized that it would be suitable for online stimulus optimization. Using an online stimulus optimization requires that stimuli be parameterized into a stimulus space in such a way that it would be meaningful to discuss how 'nearby' two stimuli are, and how to, given one good stimulus, 'move' around that stimulus to randomly generate similar but novel stimuli.

Figure 15 - Statistical summary representation of an example sound texture

A graphical summary of the statistical representation of sound textures. The original sound waveform is bandpass filtered into n_f frequency bands. The envelope of each band is then extracted, non-linearly compressed, and then summarized in several ways, as follows. **A**: Envelope marginal moments. The blue line represents the average power within each band—approximately, the power spectral density. The red line represents the variance of the envelope around its mean power. The yellow and purple lines are the raw third and fourth moments of the envelope, respectively. (The units of the y-axis are different for each moment; they are shown here scaled to the same maximum value for convenience. **B**: Modulation power. Each envelope is further modulation filtered into $\mathrm{n_{1m}}$ modulation rate bands; color indicates the total modulation power in each bin. This sound contains modulation power at mostly higher modulation rates and exhibits a weak positive correlation between envelope frequency and modulation rate. **C**: The $n_f \times n_f$ correlation matrix indicates correlation magnitude between different envelope bands. **D**: Modulation filtered spectral correlation ("C1" correlation.) Each envelope is modulation filtered into $\mathrm{n_{2m}} = 6$ modulation rate sub-bands, and then a separate between-frequency correlation matrix is constructed for each sub-band. **E**: Modulation filtered temporal modulation ("C2" correlations.) These represent a form of phase correlation between adjacent modulation bands within each spectral sub-band. These are $n_f \times (n_{2m} - 1)$ complex valued matrices; the magnitudes of the real and imaginary components are shown separately in the top and bottom rows.


A) Envelope marginal moments
Mean
Variance
3rd moment
4th moment
Envelope frequency
B) Modulation power
Modulation rate
Envelope frequency
C) Spectral correlation
Env. freq.
Env. freq.
D) Modulation filtered spectral correlation
Env. freq.
Envelope frequency
E) Modulation filtered temporal correlation
Modulation frequency
Envelope frequency

5.2 The auditory texture model

We followed as closely as possible the methods of McDermott & Simoncelli, 2011. A visual summary of the texture model is shown for one example sound in Figure 15. The texture representation of a sound comprised a small number of statistical components (summarized in Table 1): the first four marginal moments of each frequency sub-band envelope, the modulation spectra of each envelope (represented at either a fine or coarse scale), and several different correlations between envelopes and envelope modulation sub-bands. (denoted as 'C', 'C1', and 'C2' correlations in McDermott & Simoncelli, 2011).

The overall dimensionality of our texture decomposition is listed in Table 2. We used fixed dimensions $n_f = 32$, $n_{1m} = 20$, and $n_{2m} = 6$ $\mathrm{n_{1m}} = 20$as per the original McDermott & Simoncelli 2011 model. Simply adding up the total dimensionality gives a total of 8448 (although this does overestimate the 'true' dimensionality since e.g. all correlation matrices are symmetric.)

Table 1 - Summary of the statistical components in the auditory texture mode.

The dimensionality of the C2 correlations include an additional factor of two to account for their being complex-valued; the other components are all real-valued.

Name	**Notation**	**Dimensionality**
Power	$\boldsymbol{P(f)}$	$\boldsymbol{1 \times n_f}$
Variance	$\boldsymbol{V(f)}$	$\boldsymbol{1 \times n_f}$
Third moment	$\boldsymbol{M_3(f)}$	$\boldsymbol{1 \times n_f}$
Fourth moment	$\boldsymbol{M_4(f)}$	$\boldsymbol{1 \times n_f}$
Fine modulation spectrum	$\boldsymbol{M_1(f,m)}$	$\boldsymbol{n_f \times n_{1m}}$
Coarse modulation spectrum	$\boldsymbol{M_2(f,m)}$	$\boldsymbol{n_f \times n_{2m}}$
Envelope covariance (C)	$\boldsymbol{S}$	$\boldsymbol{n_f \times n_f}$
Modulation filtered envelope covariance (C1)	$\boldsymbol{S1_m}$	$\boldsymbol{n_f \times n_f \times n_{2m}}$
Modulation filtered temporal covariance (C2)	$\boldsymbol{S2}$	$\boldsymbol{n_f \times (n_{2m} - 1) \times 2}$

Table 2 - Fixed parameters for texture synthesis dimensionality

Parameter	**Description**	**Value**
n_f	Number of frequency bands	32
n_{1m}	Number of modulation bands (fine)	20
n_{2m}	Number of modulation bands (coarse)	6

5.2.1 Power spectrum

Through visual inspection of real-world auditory textures, we found that most sounds exhibited broad, unimodal-like power spectra, and could then represented as a quasi-Gaussian function as follows:

We represent a generalized gaussian function as

$$g(x|\, A, \mu, \sigma, \rho) = A\, exp\left[-\left|\frac{x-\mu}{\sigma}\right|^{\rho}\right] \tag{8}$$

where A is a gain parameter, μ is a location parameter, σ is a width parameter, and ρ is a shape parameter. This function is symmetric about μ; to support asymmetric distributions (e.g., for low-pass or high-pass spectral shapes) we make two generalized functions stepwise continuous to form a mixed generalized gaussian function:

$$h(x|\, A, \mu, \sigma_0, \rho_0, \sigma_1, \rho_1) = \begin{cases} g(x|\, A, \mu, \sigma_0, \rho_0), & x < \mu \\ g(x|\, A, \mu, \sigma_1, \rho_1), & x \geq \mu \end{cases} \tag{9}$$

where σ_0 and ρ_0 are the shape parameters for the lower side of the function, and σ_1 and ρ_1 are the corresponding parameters for the high side of the spectrum. $h(x)$ is a generic function; we denote the specific power spectrum as

$$P(f) = h(f|A_f, \mu_f, \sigma_{f0}, \rho_{f0}, \sigma_{f1}, \sigma_{f2}) \tag{10}$$

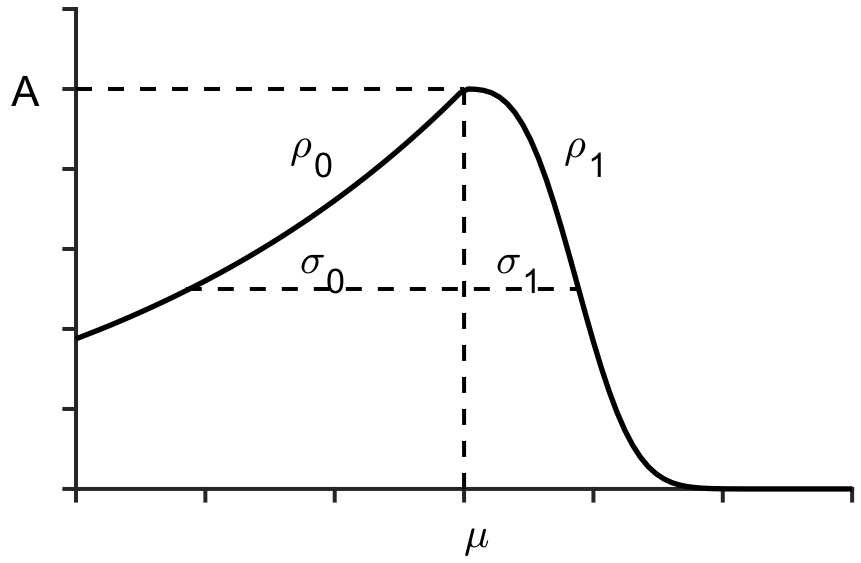


Figure 16 - Mixed generalized gaussian function

The mixed generalized gaussian function described in equation (9) has six parameters: a location parameter μ, a scale parameter A, two width parameters σ_0 and σ_1, and two shape parameters ρ_0 and ρ_1.

5.2.2 Variance spectrum

Visual inspection of real-world sound textures suggested that the amount of variance in each spectral band was related to the overall power, and this relationship was consistent between different spectral bands. (See Figure 17.) The variance spectrum was therefore parametrized with a single value $0 < \gamma_{var} < 1$:

$$V(x) = \gamma_{var} \cdot P^2(x) \tag{11}$$

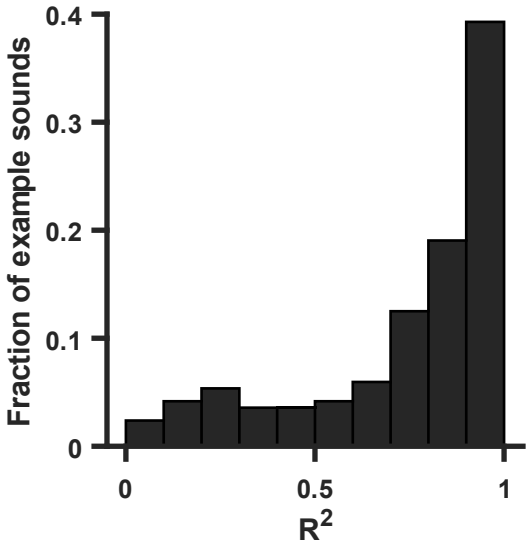


Figure 17 - Distribution of R2 values for the relationship between envelope variance and power

Each sound texture was fit to the simple model described in equation (9)

5.2.3 Modulation spectrum

The modulation spectrum of a sound can be represented as the two-dimensional function $M(f,m)$, where M is the power in the spectral bin f and the modulation bin m. The original texture specification uses two sets of modulation filters, one narrow, and one broad. Where necessary, we distinguish them with $M_1(f,m)$ and $M_2(f,m)$. Since M_2 is effectively just a down-sampled version of M_1, we have

$$\iint M_2(f,m)\, df\, dm = \gamma_{m21} \iint M_1(f,m)\, df\, dm \tag{12}$$

where γ_{m21} is a nuisance parameter that relates the total power in the two spectra.

We also have the total power constraint

$$\int M_1(f,m)\, dm = \gamma_{mod} V(f) \tag{13}$$

i.e., integrating the modulation power across all modulation bands yields the total modulation power, where γ_{mod} is another nuisance parameter. γ_{mod} is not exactly 1 because there may exist variance in the original signal falling outside the chosen modulation bands.

This still leaves the function $M_1(f,m)$ under-constrained. To simplify matters, we first specify the marginal modulation spectrum $M_{1m}(m)$ such that

$$M_{1m}(m) = \int M_1(f,m)\, df \tag{14}$$

We represent the 1D modulation spectrum $M_{1m}(m)$ in a similar manner as the power spectrum $P(f)$, with a mixed generalized gaussian. (Note that because of equation (14), the intensity parameter A_m is already effectively specified and is not included as one of the texture parameters.)

$$M_{1m}(m) = h(f|A_m, \mu_m, \sigma_{m0}, \rho_{m0}, \sigma_{m1}, \sigma_{m2}) \quad (15)$$

We have thereby have a summary of the 2D joint distribution $M_1(f, m)$ by its one-dimensional marginal distributions $P(f)$ and $M_{1m}(m)$. By Sklar's theorem, any multivariate joint distribution can be decomposed into its univariate marginal distributions and a copula function that independently describes the correlation structure between variables. We chose to use a t copula, which is specified by two parameters, the correlation coefficient $-1 < \theta_m < 1$ and degrees of freedom $v_m > 0$. The correlation coefficient θ_m has a simple interpretation: positive values mean that bands of high spectral frequencies tend to exhibit fast modulations, and low spectral bands exhibit slow modulations; negative values of θ_m mean the converse: fast modulations in low frequencies and slow modulations in high frequencies. The degrees of freedom parameter v_m affects the distribution of tail values around the mean trends in these correlations.

5.2.4 Envelope covariance

The C1 correlation coefficients in the original texture model specify envelope correlation matrices in different modulation bands, so we need to construct correlation matrices R_m for $i = 1, \ldots, n_{mod2}$. To construct general correlation matrices while

keeping the dimensionality of the parametrization low, we assume all correlation matrices are rank 1, so that we can write a correlation matrix as the outer product of a one-dimensional vector with itself. In general terms for column vector r, where $0 \leq r_i \leq 1$, we can construct a correlation matrix R with

$$R = rr' \tag{16}$$

Note that in addition to reducing the dimensionality of the parametrization, this also ensures that the correlation matrices are positive definite and symmetric, as is required for all correlation matrices.

We write the one-dimensional distribution of correlation as the mixed generalized gaussian

$$r_f(f) = h(f \mid r_{max}, \mu_r, \sigma_{r0}, \rho_{r0}, \sigma_{r1}, \rho_{r1}) \tag{17}$$

To avoid wasting time mutating correlation coefficients where little spectral or modulation power exists, we define the $r_f(f)$ shape parameters in terms of the power spectrum parameters:

$$\begin{aligned} \mu_r &= \Delta\mu_r \cdot \mu_f \\ \sigma_{r0} &= \gamma_{\sigma r} \cdot \sigma_{f0} \\ \rho_{r0} &= \gamma_{\rho r} \cdot \rho_{f0} \\ \sigma_{r1} &= \gamma_{\sigma r} \cdot \sigma_{f1} \\ \rho_{r1} &= \gamma_{\rho r} \cdot \rho_{f1} \end{aligned} \tag{18}$$

so that $r_f(f)$ is specified in terms of the parameters $\Delta\mu_r$, $\gamma_{\sigma r}$, and $\gamma_{\rho r}$.

We also assume that spectral correlation matrices in adjacent modulation bands are similar, as found by visual inspection. We then choose a one-dimensional distribution of correlation over modulation bands:

$$r_m(m) = g_{VM}(m|\mu_m + \Delta\mu_m, \kappa_{mc1}) \tag{19}$$

where $g_{VM}(x|\mu,\kappa)$ is the von Mises density function. In a similar manner to the preceding section, we construct the joint distribution $r_{fm}(f,m)$ by its marginals $r_m(m)$ and $r_f(f)$, and another t Copula with parameters θ_{corr} and ν_{corr}. Given this joint distribution, we construct the correlation matrices similarly to equation (16):

$$R_m = r_{fm}{r_{fm}}' \tag{20}$$

These correlation matrices are converted to covariance matrices with

$$S1_m = r_{max} D_m R_m D_m \tag{21}$$

where r_{max} is a scalar parameter specifying the maximum correlation value overall; and D_m are diagonal matrices with diagonals equal to the total variance in each envelope.

The next step is to construct the overall envelope covariance matrix S, i.e. the covariance matrix between envelopes before any modulation filtering. Clearly, this matrix must be related to the individual modulation-filtered covariance matrices S_{1m}. Based on visual inspection of real-world textures we choose

$$(S)_{i,j} = \begin{cases} \sum_m (S1_m)_{i,j} \quad , & i = j \\ \lambda_{corr} \sum_m (S1_m)_{i,j} \quad , & i \neq j \end{cases} \tag{22}$$

where $0 \leq \lambda_{corr} \leq 1$; i.e. the diagonals (the variances) sum linearly, to form the total variance in each envelope, but off-diagonals (covariances) sum sub-additively with rate λ.

5.2.5 Modulation covariance

The C2 correlation structures were the most difficult to parametrize. Their high dimensionality and non-standard definition made them difficult to visualize and interrogate. Initial attempts to synthesize arbitrary C2 values ran into poor convergence, which might have been caused by interactions between C1 and C2 correlations: C1 correlations define how similar two different envelopes are; C2 correlations define the temporal structure within an envelope; thus, if two envelopes have a high C1 correlation, they must also have similar C2 correlations. Therefore, synthetizing arbitrary C2 values requires matching a constraint specified by the C1 correlations, as discussed below.

We indicate the complex valued C2 covariance matrices as $S2_{f,m}$ for frequency band f and modulation band $m = 1, \ldots, n_{mod2} - 1$. We summarize C2 values by summing their values over all frequency and modulation bands to yield the scalar complex value $\widehat{S2}$:

$$\widehat{S2} = \sum_{f,m} S2_{f,m} \tag{23}$$

Since $\widehat{S2}$ is complex, we can represent it by its phase $\phi_{c2} = \angle\widehat{S2}$ and magnitude $A_{c2} = \left|\widehat{S2}\right|$. Note that A_{c2} can be small if either the individual components of C2 are

small, or if components are large but cancel each other out. To distinguish these possibilities, we also specify the total C2 power fraction:

$$\gamma_{c2} = \frac{\sum_{f,m}|S2_{f,m}|}{\iint M_2(f,m)\, df\, dm} \tag{24}$$

To address these constraints, we define the average C1 correlations between adjacent modulation bands as

$$T1_m = \frac{C1_m + C1_{m+1}}{2} \tag{25}$$

where $m = 1, \ldots, n_{mod2} - 1$.

We also define C2 similarity matrices by first quantifying how similar C2 covariances in different frequency bands are as:

$$similarity_{m,i,j} = 1 - 2\frac{|S2_{i,m} - S2_{j,m}|}{|S2_{i,m}||S2_{j,m}|} \tag{26}$$

and then mapping that value to the range [-1, 1] with

$$(T2_m)_{i,j} = \alpha_{c2}\, tanh\left[similarity_{m,i,j} + \beta_{c2}\right] \tag{27}$$

This allows $-1 \leq (T2_m)_{i,j} \leq 1$ to quantify how similar the C2 covariances are between spectral bands i and j for modulation band m. α_{c2} and β_{c2} are arbitrary scalar parameters that control how C2 similarities map to C1 correlations. Note that $T2_m$ is effectively a correlation matrix and has the same dimensions as the C1 correlation matrices $T1_m$. Instead of comparing $T1_m$ and $T2_m$ directly, which could introduce

numerical issues by comparing correlations between very weak signals, we use the total modulation power to convert correlations to covariances, i.e.

$$U1_m = E_m T1_m E_m \tag{28}$$

and

$$U2_m = E_m T2_m E_m \tag{29}$$

where E_m is a $n_f \times n_f$ diagonal matrix with diagonals

$$(E_m)_{f,f} = \sqrt{M(f,m) \cdot M(f,m+1)} \tag{30}$$

where $M(f,m)$ is, as previously defined, the total power in spectral band f and modulation band m.

Now, we define the C1/C2 similarity error as

$$Err_{sim} = \sum_{m,\ i<j} \left((U1_m)_{i,j} - (U2_m)_{i,j}\right)^2 \tag{31}$$

the direction error as

$$Err_{dir} = \left| \left(\sum_{f,m} S2 \right) - \widehat{S2}_{target} \right|^2 \tag{32}$$

and the power error as

$$Err_{pwr} = \left(\sum_{f,m} \left| S2_{f,m} \right| - \gamma_{c2} \iint M_2(f,m)\, df\, dm \right)^2 \tag{33}$$

The total error is then

$$Err_{total} = Err_{sim} + Err_{dir} + Err_{pwr} \tag{34}$$

We can then find a set of C2 correlations that minimize Err_{total}, under the constraints of the predetermined C1 correlations, and the target summary statistic $\widehat{S2}_{target}$, by using standard bounded gradient descent methods, specifically the Matlab function *fmincon*.

The bounds for this optimization are as follows: Consider the C2 covariances $S2_{f,m}$, $f = 1, \dots, n_f$ and $m = 1, \dots, n_{2m} - 1$. The modulation dimensionality is $n_{2m} - 1$ because the C2 covariances are defined only between adjacent modulation bands, and not all pairs of modulation bands like a standard covariance. However, if we treated it like a standard covariance, we could write covariance matrices S'_f of size $n_{2m} \times n_{2m}$ for $f = 1, \dots, n_{2m}$. In that case, we would have

$\left(S'_f\right)_{m,m}$	$= M_2(m, f)$	since the diagonals of a covariance matrix are the variances,
$\left(S'_f\right)_{m,m+1}$	$= S2_{f,m}$	the C2 covariance defined between adjacent modulation bands,
$\left(S'_f\right)_{m+1,m}$	$= S2^*_{f,m}$	since the covariance matrix must be conjugate symmetric, and
$\left(S'_f\right)_{m,n}$	$= 0$	for all non-adjacent modulation bands.

This is to say that the C2 covariance matrix S'_f is a conjugate symmetric tri-diagonal matrix with variances on the main diagonal and C2 covariances on the first upper and lower diagonals. The bounding condition for finding $S2$ under the previously described objective function are that S'_f must remain valid covariance matrices, i.e. S'_f are positive semi-definite, or

$$\det(S_f') \geq 0, f = 1, \ldots, n_f \tag{35}$$

5.2.6 Higher order moments

Given a set of C2 correlations, we are ready to then specify the third and fourth moments. We denote the moments of each spectral band f as $M_3(f)$ and $M_4(f)$. These values are necessarily constrained by the C2 correlations because the C2 correlations partially characterize temporal transients in the envelope, as illustrated in Figure 18. Instead of specifying the raw moments directly, we specify them in a normalized way with a variance normalized version of the third moment known as skewness:

$$skewness(f) = \frac{M_3(f)}{\big(V(f)\big)^{3/2}} \tag{36}$$

where $V(f)$ is the variance of each spectral band. The equivalent normalized version of the fourth moment is the bimodality coefficient (BMC):

$$BMC(f) = \frac{\big(skewness(f)\big)^2 + 1}{M_4(f)}\big(V(f)\big)^2 \tag{37}$$

As can be seen in 5, the real and imaginary components of the C2 correlation put upper and lower bounds on the allowable range of the skewness and BMC. We denote the normalized position within this 5-95th percentile range as $0 \leq \gamma_{skew} \leq 1$ and $0 \leq \gamma_{BMC} \leq 1$.

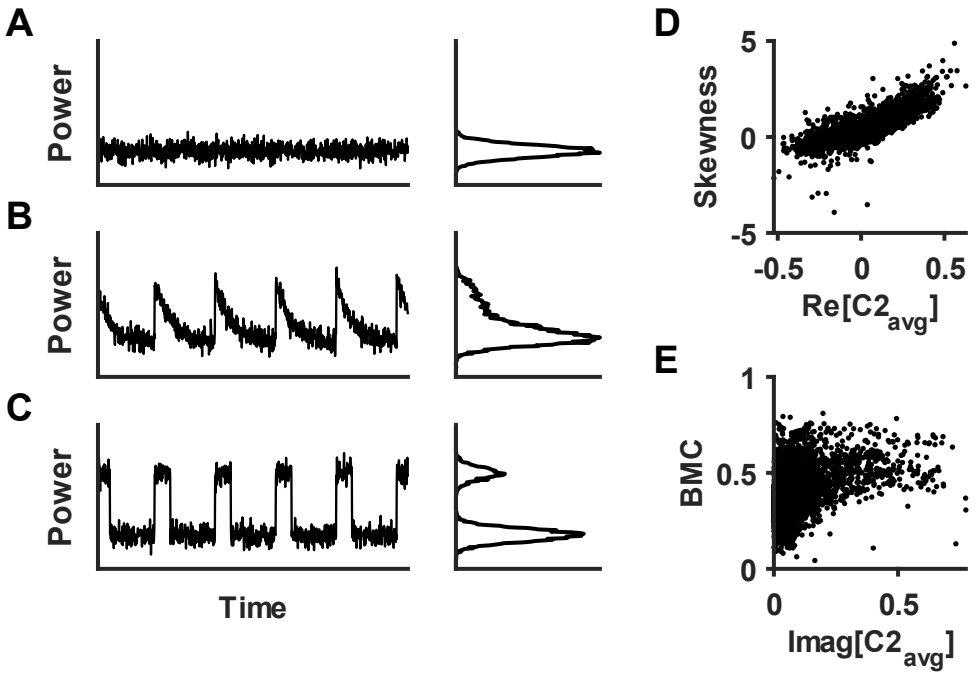


Figure 18 - Relationship between envelope marginals and C2 correlations

A-C: Example cartoon envelopes (left) and their marginal distributions (right). **A**: A gaussian distributed envelope. **B**: An envelope with onset transients. Its marginal distribution exhibits positive skewness. **C**: An envelope with onset and offset transients. Its marginal exhibits bimodality. **D**: Observed correlation between envelope skewness and the real component of the C2 correlation statistic in the real-world sound bank, averaged across temporal bands. **E**: Observed correlation between the envelope bimodality coefficient (BMC) and the imaginary component of the average C2 statistics

Table 3 - Summary of synthetic texture parameter space

Category	Variable	Description	Initial range
Power spectrum	μ_f	Peak frequency	1 oct. centered on BF
	A_f	Peak intensity	0.9 – 1.1
	σ_{f0}	Lower size parameter	1 – 5 bins
	ρ_{f0}	Lower shape parameter	.5 – 5
	σ_{f1}	Upper size parameter	1 – 5
	ρ_{f1}	Upper shape parameter	.5 – 5
Variance spectrum	γ_{var}	Total modulation depth	10^{-5} – 0.1
Modulation spectrum	γ_{m21}	Ratio of total power in the two different modulation filter sets	.4 – .6
	γ_{mod}	Fraction of total modulation power in modulation spectrum	
	μ_m	Peak modulation rate	1 – n_{mod}
	σ_{m0}	Lower size parameter	1 – 10
	ρ_{m0}	Lower shape parameter	.5 – 5
	σ_{m1}	Upper size parameter	1 – 10
	ρ_{m1}	Upper shape parameter	.5 – 5
	θ_m	Modulation spectrum copula correlation coefficient	-0.9 – 0.9
	ν_m	Modulation spectrum copula degrees of freedom	.5 – 10
Envelope covariance	r_{max}	Maximum correlation	.01 – 1
	$\Delta\mu_r$	Correlation peak offset from spectral peak	-0.75 – 0.75
	$\gamma_{\sigma r}$	Width scaling, relative to spectral width parameters	0.6 – 1.4
	$\gamma_{\rho r}$	Shape scaling, relative to spectral shape parameters	0.6 – 1.4
	$\Delta\mu_m$	Correlation peak offset from modulation peak	-0.75 – 0.75
	κ_{mc1}	Variance of the modulation correlation distribution	-0.75 – 0.75
	θ_{corr}	Correlation/modulation copula function correlation coefficient	0.8 – 1.2
	ν_{corr}	Correlation/modulation copula function degrees of freedom	0.5 – 10
	λ_{corr}	C1 to C covariance additivity	0.1 – 1
Modulation covariance	γ_{c2}	Fraction of total modulation power	0.5 – 0.75
	ϕ_{c2}	Average C2 covariance angle	0 – $\frac{\pi}{2}$
	A_{c2}	Average C2 covariance magnitude	10^{-4} – 0.1

	α_{c2}	C2 similarity to C1 correlation mapping scale parameter	-0.9 – 0
	β_{c2}	C2 similarity to C1 correlation mapping location parameter	1 – 10
Higher order moments	γ_{skew}	Normalized skewness	0 – 1
	γ_{BMC}	Normalized bimodality	0 – 1

5.2.7 Summary of the synthetic texture specification algorithm:

1. Specify power spectrum parameters: μ_f, A_f, σ_{f0}, ρ_{f0}, σ_{f1}, ρ_{f1} to obtain $P(f)$
2. Specify modulation power with the total modulation depth γ_{var} to obtain $V(f)$
3. Specify fraction of modulation in modulation filter bank γ_{mod} and the marginal modulation spectrum parameters μ_m, σ_{m0}, ρ_{m0}, σ_{m1}, ρ_{m1} to obtain $M_{1m}(m)$
4. Specify t Copula parameters θ_m and ν_m to combine $V(f)$ and $M_{1m}(m)$ to yield $M_1(f,m)$
5. Specify $\Delta\mu_r$, $\gamma_{\sigma r}$, $\gamma_{\rho r}$ to obtain marginal spectral correlation distribution $r_f(f)$
6. Specify $\Delta\mu_m$, κ_{mc1} to obtain marginal modulation correlation distribution $r_m(m)$
7. Specify t Copula parameters θ_{corr} and ν_{corr} to obtain $r_{fm}(f,m)$
8. Specify r_{max} and use r_{fm} to obtain C1 covariance matrices $S1_m$
9. Specify λ_{corr} and use R_m to obtain C covariance matrix S
10. Specify C2 target parameters γ_{c2}, ϕ_{c2}, A_{c2}; C2 similarity to C1 covariance mapping parameters α_{c2}, β_{c2}, and use gradient descent to find a set of C2 covariance matrices $S2$
11. Specify γ_{skew} and γ_{BMC} and use C2 correlations to obtain third and fourth moments $M_3(f)$ and $M_4(f)$

5.3 Results

As mentioned in the introduction, a major hurdle to overcome was how to represent synthetic textures in a parametric way that made an evolutionary approach feasible. The example texture shown in Figure 15 exhibits 8448 different values, a dimensionality much too large to optimize productively. The first approach we attempted to reduce this dimensionality was to take a collection of real-world sounds, measure their texture, apply PCA to that set of summary statistics, and represent synthetic textures by their projections on to the first few principle components. We found that the gradient descent method used to synthesize sounds from their texture often failed to converge when applied to textures generated by this approach. This presumably occurred because PCA only captures linear relationships between variables, but the texture specification includes many complicated non-linear relationships between variables. Ignoring these relationships would lead to the specification of a set of summary statistics for which a solution (in terms of a sound waveform) could not exist.

In lieu of an out-of-the box solution, we therefore performed a manual parametrization of auditory textures that might account for the major non-linear constraints identified in the texture model. This parametrization was not designed to account for all possible textures (for example, we assume a unimodal like distribution of spectral power, which necessary precludes spectra with two or more distinct peaks) but rather span a large enough space to encompass both higher level neural receptive fields and a wide array of statistical structures.

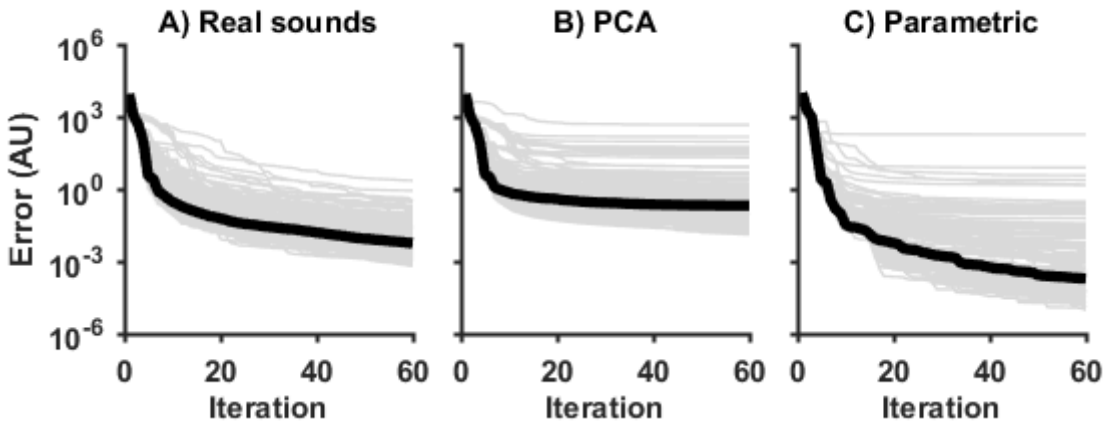


Figure 19 – Texture synthesis gradient descent results

Each thin grey line is the error trace of a single run of the gradient descent sound synthesis for a different set of target statistics (n = 168). The thick black line shows the median over all sounds. **A**: Each trace represents a single run of a synthesis of a sound with target statistics taken from real world example sounds. Most traces follow a stereotyped decrease in error with each iteration. **B**: Target statistics were randomly generated points in a space defined by the first 32 principle components of the sounds used in A. The decrease in error rapidly plateaus. (Varying the number of components used did not meaningfully affect this pattern.) **C**: Target textures were random points defined by the texture parametrization used in this manuscript. Although some error traces plateau, the majority behave like the pattern observed in A. The traces that plateau very early tended to result in waveforms that could not be played and would be discarded in the evolutionary optimization. (See methods.)

Figure 19 shows a summary and comparison of the two different synthetic texture synthesis approaches for a variety of example sounds. In Figure 19B, random sounds constructed from randomly choosing points in the principle component space tend to plateau early, as described above, whereas random sounds using the parametric approach behave more similarly to sounds constructed from observed real world statistics.

5.4 Conclusions

Through a process of visual inspection of the texture representation of real-world sounds, we were able to develop a 31-dimensional parametrization of the texture model that allowed for randomly specified textures to be synthesized with an error pattern that was similar to those of real-world sounds. This was a necessary precondition to do online stimulus optimization as described in the following chapter.

6. Neural Representation of Auditory Texture Statistics in Non-Primary Auditory Cortex

6.1 Introduction

After having developed a method in the previous chapter to represent synthetic auditory textures in a low dimensional space, we were ready to use this parametrization in an online evolutionary optimization algorithm, where stimuli that evoked a weak response could be modified and improved, allowing experimental time to be concentrated around a neuron's receptive field. Evolutionary stimulus optimization using genetic algorithms are a suitable approach for this problem, having been used in the past to optimize high-dimensional stimuli in extra-striate visual cortex, (Carlson, Rasquinha, Zhang, & Connor, 2011; Hung, Carlson, & Connor, 2012; Vaziri, Carlson, Wang, & Connor, 2014; Vaziri & Connor, 2016; Yamane, Carlson, Bowman, Wang, & Connor, 2008), including visual texture, (Okazawa, Tajima, & Komatsu, 2015), as well as low-level stimulus features in primary auditory cortex. (Chambers, Hancock, Sen, & Polley, 2014). We also developed a custom algorithm to synthesize stimuli in a manner that was fast enough to use in an online setting. We applied our online synthetic texture optimization to well-isolated single neurons in the non-primary auditory cortex of marmoset monkeys in a passive listening condition. For each neuron, we ran two independent lineages of the optimization analysis in parallel and fit a model to data

from each lineage that could successfully predict firing rate responses to independent stimuli in the alternate lineage. To characterize how well these models represented different components of the auditory texture structure, we combined the modeled neurons together in a single virtual population and demonstrated how it could explain several of the patterns observed in human discrimination of auditory textures.

6.2 Methods

6.3 Evolutionary optimization algorithm

6.3.1 Stimulus synthesis

The McDermott and Simoncelli (2011) algorithm for synthesizing auditory textures was rather slow, (approximately ~40 minutes per 5 second stimulus), which was not suitable for online stimulus synthesis for neurophysiological recording. This was due to two reasons: 1) a lot of computational time is spent each iteration on filtering the noise into its component sub-bands and extracting their envelopes, and 2) the gradient descent is performed using a numerical steepest descent method, which requires a lot of function evaluations to estimate the gradients.

We therefore developed a novel texture synthesis approach that would be fast enough to use in an online context. The first thing we did to support faster synthesis times was to assume that we could operate directly on the sub-band envelopes, and covert them to waveforms by up-sampling and multiplying by appropriate carriers. This conversion to waveform needs to occur only once, after finishing the gradient descent envelope optimization. A second advantage to this approach was that since the envelopes are necessarily down-sampled relative to the original waveform, the envelopes require smaller memory footprint, again making each iteration of the gradient descent algorithm more efficient. The third and final advantage to operating directly on the envelopes was that it made plausible the analytical expression of the cost function and

its Jacobian; this made it possible to perform a local linearization of the error function and use a conjugate gradient descent method to perform more efficient and robust gradient descent.

The cost function between the observed texture statistics T^{obs} and the target texture T^* was defined as the weighted sum of the error terms for each statistical component.

$$\begin{aligned}&\text{Cost}(T^{obs}, T^*)\\&= \alpha_{M1}E_{M1} + \alpha_{M2}E_{M2} + \alpha_{M3}E_{M3} + \alpha_{M4}E_{M4} + \alpha_{mod}E_{mod} + \alpha_{C1}E_{c1} + \alpha_{C2}E_{C2}\end{aligned} \tag{38}$$

Where the α terms are scalar non-negative weights, and the E terms are the errors in each statistical component. E_{M1} is the sum of squared errors in the envelope means, e.g.

$$E_{M1} = \sum_{i=1}^{n_f} \left(P^o(i) - P^*(i)\right)^2 \tag{39}$$

Where n_f is the number of frequency bins, and $P_f^o(i)$ and $P_f^*(i)$ are the envelope means in the i-th frequency bin of the observed and target texture statistics E_{M2}, E_{M3}, and E_{M4} are defined similarly using the envelope variance $V(i)$, third moments $M_3(i)$, and fourth moments $M_4(i)$. The error in the modulation spectrum is summed over all spectrotemporal bins:

$$E_{mod} = \sum_{i=1}^{n_f} \sum_{j=1}^{n_{m1}} \left(M_1^{obs}(i,j) - M_1^*(i,j)\right)^2 \tag{40}$$

and the errors in the C1 and C2 covariances are defined similarly.

Waveform synthesis proceeded as follows:

1. An initial guess for the envelopes was generated by drawing samples independently from a folded normal distribution.
2. Conjugate gradient descent was run for 60 iterations to minimize the error between the target texture statistics and the observed statistics of the envelopes
3. The envelopes were converted to a single temporal wave form by upsampling the envelopes and multiplying by a carrier with the same power spectrum as the correspond frequency band
4. The texture statistics of this temporal waveform were measured and used for subsequent data modeling and other texture control stimuli (see below)

Ultimately, our approach could synthesize sounds in approximately three minutes, which made online synthesis feasible. Since each stimulus takes on the order of three minutes to synthesize, all stimuli in each generation were generated in parallel so that the total time to synthesize multiple stimuli was equivalent to synthesizing a single stimulus. Custom software in Matlab and Python transferred texture specifications to a high-performance computing cluster (Maryland Advanced Research Computing Center, MARCC), monitored the synthesis progress, and then transferred the synthesized waveforms back to the experimental computer. Occasionally, waveforms with very large peak intensities were generated, usually due to failure of the gradient descent to converge, yielding a waveform that was too loud to be played at full dynamic range. These stimuli were simply discarded without playing or contributing further to

the evolutionary optimization. All other synthesized sounds were used in the online experiment, regardless of the gradient descent success.

Evolutionary optimization proceeded along two independent lineages; while stimuli from one lineage were being presented, stimuli for the next generation of the other lineage were being synthesized. Generation sizes were chosen to approximately balance presentation and synthesis time to maximize efficiency. Since the computing cluster is a shared resource, occasionally under high load we encountered longer wait times and generation sizes were adjusted to account for this. In general, the number of stimuli generated varied from 24-32 stimuli.

Sounds were 500 ms long, played with a 1500 ms SOA. Sounds were played for a minimum of 5 repetitions. In cases when the synthesis for the next generation was delayed, this was occasionally increased to 6 or 7 repetitions to avoid downtime while waiting for the next generation. The spontaneous rate was calculated separately for each generation from the 500 ms pre-stimulus period collapsed across all stimuli. Responses were measured from the time window 500-1100 ms.

6.3.2 Initial generation

Initial characterization of the neuron began with traditional stimulus sets of tones and noise, where we estimated a neuron's preferred speaker location and intensity as well as its preferred frequency. The initial distribution of the center frequency μ_f was then tailored to the neuron's receptive field: the initial distribution was uniformly distributed

(in log-frequency space) in a one octave range centered around the neuron's preferred frequency. Additionally, although the distribution of the intensity parameter A_0 was not changed for each neuron, a reference intensity was chosen to place most stimulus attenuations near the preferred intensity of the sound. Subsequent mutation of the center frequency and intensity parameters were not clipped to these ranges in case these initial estimates were incorrect. The distributions of the remaining parameters were the same for all neurons, as listed in Table 3.

Most parameter values were generated independently, except for each pair of generalized gaussian shape parameters σ_0 and σ_1 for the spectral, modulation, and correlation distributions, which were always jointly coupled with a bivariate Frank copula and logarithmically uniform marginal distributions. This meant the most common random draw was with both σ_0 and σ_1 small; the least common draw was with both σ_0 and σ_1 large. The rationale behind this was to bias the distribution towards more frequent narrow spectra and less frequent broad spectra, since narrower spectra would require denser sampling.

6.3.3 Subsequent generations

The breeding pool for synthesizing a new generation was all stimuli from the preceding three generations of the same lineage that evoked a significant response, with some exceptions described below. Stimuli for the new generation belonged to one of the following classes:

Parameter elites: The top 3 highest-response stimuli from the breeding pool were selected to yield parameter elites: stimuli whose parameter sets were reused, without mutation. Because of the way the C2 and third and fourth moments were constructed from their parameters, this meant that parameter elites resulted in non-identical texture specifications.

Texture elites: The single highest-response stimulus from the breeding pool was selected to yield a texture elite, where instead of reusing the same parameter set, we reuse the actual texture specification. Since there is some stochasticity in the generation of a texture from a parameter set, this allows us to more reproducibly test similar locations in texture space.

Texture controls: The same texture used for the texture elite described above was used to synthesize alternative sounds with a perturbed synthesis method, to test mutations in a way that could not be directly captured by mutations in parameter space. This was done in one of three different ways:

1. Constraints on envelope correlations were removed
2. Constraints on C2 modulation correlations were removed
3. Constraints on the third and fourth moments were removed

Because texture controls could not be described in the normal parameter space, they were not eligible for subsequent mutation and were thus ineligible for the breeding pool.

Mutations: Stimuli were chosen randomly, without replacement, from the breeding pool. Random selection was weighted by evoked responses, so that higher-response stimuli were more likely to be selected. Mutation occurred by taking independent random steps in all 31-dimensions simultaneously. Each random step was drawn from a t-distribution with five degrees of freedom and mean zero. Standard deviation of each parameter's mutation was 10% of its initial range, as listed in table X. The standard deviations of the location parameters μ_f and μ_m were further multiplied by min $[\sigma_{f0}, \sigma_{f1}]$ and min $[\sigma_{m0}, \sigma_{m1}]$ respectively, so that broad spectra took larger steps in center frequency than narrow ones.

Several other specialized 'global' mutations occurred with a 10% chance on each mutation, as follows:

- Spectrum shape flipping: The lower (σ_0, ρ_0) and upper (σ_1, ρ_1) shape parameters swap values. (This could, for example, turn a high-pass shape into a low-pass one.) This occurred independently for both the frequency and modulation spectra.
- Spectral/temporal correlation flipping: The spectrotemporal correlation coefficient θ_m would flip sign.
- Correlation offsets: The correlation distribution offsets $\Delta\mu_r$ and $\Delta\mu_m$ would flip sign independently.

In general, all mutation rates were scaled by firing rate so that higher-response stimuli mutated with smaller steps, to more densely sample the center of the receptive field.

Novel stimuli: To increase genetic diversity and prevent getting stuck in local minima, a minimum of two stimuli per generation were generated from the original distributions specified for the initial generation. In the case where the breeding pool was too small to produce enough elites or mutations, more than two novel stimuli were created, up to the specified generation size.

6.4 Data analysis

6.4.1 Dataset

Spontaneous firing rates were calculated from the pre-stimulus time period of 0 – 500 ms. Average stimulus evoked firing rates were calculated from a time window of 520 – 1100 ms. We included all neurons for which we recorded at least five generations in both lineages, and at least one stimulus in either lineage evoked a firing rate more than 20 Hz above the spontaneous rate.

6.4.2 Basic analysis of the evolutionary optimization

For each neuron, we fit a three-way ANOVA model on the average stimulus evoked firing rate with factors stimulus origin (‘new’, ‘mutated’, or ‘elite’), lineage (1 or 2), and generation using the Matlab function *anovan*.

6.4.3 Texture receptive field

For each neuron, we constructed a 'texture receptive field' (TRF) as a first order linear model:

$$\bar{r}_i = \beta T_i + \epsilon \tag{41}$$

Where $\bar{r}_i$ is the average firing rate in response to texture T_i, β is a vector of coefficients, and ϵ is i.i.d. Gaussian noise. Each texture T_i was represented as a n_p-element vector of all its statistical summary values, where n_p = 4560 is the total number of parameters in the full texture model. Including the constant term, there were therefore 4561 β coefficients to estimate. Since each neuron had only on the order of 100-300 observations, this model is highly under-constrained. To handle this, we used elastic net regularization. (Hastie, Tibshirani, & Friedman, 2009) Briefly, this technique combines L_1 and L_2 regularization to find

$$\hat{\beta} = \underset{\beta}{argmin}[\|\bar{r} - \beta T\|^2 + \lambda_2\|\beta\|^2 + \lambda_1\|\beta\|] \tag{42}$$

Where λ_1 and λ_2 are the L_1 and L_2 weights, respectively. Elastic net regularization is particularly suited to cases where 1) the solution is expected to be sparse (i.e. most elements of β are zero) and 2) many independent variables are correlated (since the texture model exhibits a lot of autocorrelation, where elements nearby in frequency or modulation spectra have similar values.) To optimize λ_1 and λ_2, we use ten-fold cross-validation within each lineage to find the values of λ_1 and λ_2 that minimize the out-of-fold prediction error (MSE). Fold partitioning was done by sorting stimuli by firing rate, and then randomly assigning every tenth stimulus to a fold. This ensured that every fold contained a distribution of firing rates than spanned the smallest and largest firing

rates exhibited by that neuron. Elastic net regression was performed using the Glmnet library in Matlab (http://web.stanford.edu/~hastie/glmnet_matlab/).

Using the optimized values of λ_1 and λ_2 we fit a model to the entire dataset from one lineage and measured how well it predicted the firing rates of the stimuli from the other lineage. We define $\rho_{1\mapsto 2}$ as the correlation coefficient between observed responses to stimuli in lineage 2 and the responses predicted by a model fit to data from lineage 1, and vice versa for $\rho_{2\mapsto 1}$. We summarized the overall model efficacy with the average of the two correlation coefficients $\rho = \frac{1}{2}(\rho_{1\mapsto 2} + \rho_{2\mapsto 1})$. Finally, we normalized the estimate of ρ using the Spearman-Brown corrected split-half self-consistency as a measure of neural reliability. This was done separately for each neuron by randomly splitting the repetitions of each stimulus into two folds and measuring the correlation between the stimulus average firing rates between the two folds. We denote the corrected average correlation for each neuron as $\bar{\rho}$.

6.4.4 STRF

As an alternative to the TRF model, we fit a more traditional spectrotemporal receptive field (STRF) by binning spike times and extracting the segments of the envelopes preceding each spike. STRFs were fit by elastic net regression in the same manner as the TRFs. Responses were predicted by summing over time the convolution of the STRF and stimulus envelopes.

6.4.5 Noise modeling

For each neuron we fit a linear regression model relating its average firing rate $\bar{r}$ to its standard deviation of firing rate σ:

$$\sigma(\bar{r}) = exp[\alpha \, log \, \bar{r} + \sigma_0] \tag{43}$$

for Poisson noise, $\sigma = \bar{r}$ which would lead to fitted values $\alpha = 1$ and $\sigma_0 = 0$, but all neurons had highly sub-Poisson noise.

6.4.6 Virtual response

We define the virtual response r_i^{virt} of neuron i to the k-th repetition of texture T_j as:

$$r_{ijk}^{virt} = clamp\left[\beta_i T_j + \epsilon\left(\sigma_i(\beta_i T_j)\right)\right]_{r_i^{min}, r_i^{max}} \tag{44}$$

where $\beta_i T_j$ is the expected mean response, σ_i is the previously described noise model for neuron i, and $\epsilon(\sigma)$ is random Gaussian variable with mean zero and standard deviation σ. The virtual response r_{ijk}^{virt} thus includes both the estimated mean response, and estimated noise around that mean. The $\text{clamp}[\cdot]_{r_i^{min}, r_i^{max}}$, function clamps its argument to the range specified by the minimum and maximum observed firing rates of neuron i to prevent unrealistic firing rate extrapolations.

6.4.7 Classifier training

We obtained the n_T = 168 item collection of example sound textures (McDermott, personal communication). The statistical summary was measured for each sound and represented as a texture vector T. The one free parameter in the virtual stimulus presentation was how 'loud' to present each stimulus, i.e. the total envelope power. We

chose this by first calculating the minimum and maximum RMS power of the stimuli presented to the neuron in the online stimulus optimization experiments, and then estimating the 'best' level within this range, i.e. the power for which the TRF predicted the maximum response to the example sound texture. This best power was used for all subsequent virtual presentations of that texture, including its 'perturbed' versions. (See below.)

We generated n_{reps} = 10 virtual repetitions of each sound with equation (44) for each neuron to construct a $\left(n_{reps} \cdot n_T\right) \times n_{neurons}$ = 1680×36 design matrix of virtual responses. This design matrix and the corresponding list of sound names was used to fit a linear classifier using the *fitcdiscr* function in Matlab.

6.4.8 Classifier evaluation

We attempted to replicate as closely as possible the task design from McDermott & Simoncelli, 2011. For each trial, a texture was generated whose target statistics were specified by a corresponding real-world sound's observed texture statistics. A new synthetic sound was synthesized, with error weights determined by the class of perturbed texture. (E.g., for the 'power only' condition, all the α terms in equation (43) except for α_{M1} were set to zero, so only errors in the envelope power spectrum affected the synthesis process.) The observed statistics of the synthesized sound were then used as in the input vector T for equation (44) to yield a new virtual population response vector. This vector was then supplied to the linear classifier, along with a list of five

sound names, one of which was the original sound's name (i.e. we made a five-alternative forced choice 'task' for the linear classifier to perform.) The category that the classifier assigned the highest posterior probability was taken as the classifier's choice and scored as either correct or incorrect depending on whether that selected category matched the original name of the sound. This process was repeated 100 times for each example sound in the collection to obtain distributions of classifier performance.

6.4.9 Texture Discrimination Analysis

6.4.9.1 Pairwise texture distances

To quantify how individual neurons could distinguish between 'same' and 'different' texture categories, we classified pairs of synthetic textures based on how similar they were in texture space. We defined the texture distance ΔT_{ij} between a pair of textures T_i and T_j using the same cost function used for the gradient descent stimulus synthesis in equation (38):

$$\Delta T_{ij} = Cost(T_i, T_j) \tag{45}$$

For each neuron, we collapsed both evolutionary lineages into a single dataset of n_t synthetic textures and calculated the distances between all $n_p = n_t(n_t - 1)$ pairs of textures. We kept for analysis all pairs where the neuron responded to at least one of the pair above 50% of its maximum response. Next, the distribution of pairwise texture distances ΔT_{ij} was split into quartiles, and the middle two quartiles were dropped. The

first and last quartiles were then datasets of either relatively small or large pairwise distances; for convenience they were labeled 'similar' or 'dissimilar' pairs, respectively. More casually, pairs of textures with small ΔT values belong to the same 'type' or 'class' of texture, whereas pairs of textures with large ΔT values belong to different types or classes.

6.4.9.2 Texture type discrimination

To estimate how neural firing rates might discriminate between similar and dissimilar textures as a function of stimulus duration, we simulated changes in stimulus duration by estimating firing rates over time windows of varying duration that were less than or equal to the actual stimulus duration. For example, even though stimulus durations were fixed to 500 ms, we simulated a response to a 50 ms stimulus by randomly choosing a 50 ms time window within that stimulus and calculating the average firing rate within that same window over each repetition.

For each pair of textures, we define the pairwise firing rate difference between neural responses to those textures as a function of the time window size Δt as

$$\Delta \bar{r}_{ij}(\Delta t) = \left|\bar{r}_i(t_0, t_0 + \Delta t) - \bar{r}_j(t_1, t_1 + \Delta t)\right| \quad (46)$$

where $\bar{r}_i(a, b)$ is the firing rate of the neuron in response to stimulus i between time points a and b, averaged over stimulus repetitions, and t_0 and t_1 are randomly chosen timepoints within the stimulus presentation window. Then, given two distributions of Δr_{ij}, one for the small ΔT_{ij} pairs, and one for the large ΔT_{ij} pairs, we constructed an ROC curve for the discrimination of small and large ΔT. This discriminability was then

summarized by numerically integrating the area under the ROC curve (AUC). We define the correlation coefficient ρ_{Type} as the correlation between the AUC and the time window duration:

$$\rho_{Type} = corr\left[\Delta t, AUC_{Type}(\Delta t)\right] \quad (47)$$

where a positive value of ρ_{Type} indicates that the neurons ability to discriminate between small and large ΔT increases with increasing window duration. Finally, since the observed value of ρ_{Type} depends on the random timepoints chosen for each stimulus, we calculate bootstrap statistics by repeating this process 100 times to obtain a distribution of observed $\rho_{\text{Type}}^{\text{Obs}}$ values, and again 100 more times, randomly reshuffling the large and small datasets each iteration to obtain a distribution of $\rho_{\text{Type}}^{\text{Shuff}}$ under the null hypothesis. We reported the median value of the observed $\rho_{\text{Type}}^{\text{Obs}}$ statistic, and consider the correlation significant if the distributions of $\rho_{\text{Type}}^{\text{Obs}}$ and $\rho_{\text{Type}}^{\text{Shuff}}$ were significantly different according to a ranksum test with $p < 0.05$.

6.4.9.3 Texture token discrimination

A similar analysis was performed to quantify how well neurons could discriminate different tokens of the same texture type. In this case, we analyzed only pairs of textures that belonged to the small ΔT dataset, as described above. Instead of calculating the rate difference averaged over repetitions, we compared firing rate differences between individual presentations of the same or different stimuli.

With this approach, we define $\Delta r_{iuv}^{\text{same}}$ as the difference in firing rate between two different repetitions u and v of the *same* stimulus i:

$$\Delta r_{iuv}^{\text{same}}(\Delta t) = |r_{iu}(t_0, t_0 + \Delta t) - r_{iv}(t_0, t_0 + \Delta t)| \quad (48)$$

Note that if a neuron responded identically to every repetition of the same stimulus, Δr^{same} would always be zero.

We also define $\Delta r_{ijuv}^{\text{diff}}$ as the difference in firing rate between repetition u of stimulus i and repetition v of a different stimulus j.

$$\Delta r_{ijuv}^{\text{diff}}(\Delta t) = \left|r_{iu}(t_0, t_0 + \Delta t) - r_{jv}(t_0, t_0 + \Delta t)\right| \quad (49)$$

The timepoint t_0 in both cases is still chosen randomly as described above for the type discrimination analysis. For a neuron presented m different stimuli each with n repetitions, there would be $mn(n-1)$ values of Δr^{same} and $mn^2(m-1)$ values of Δr^{diff}. We therefore obtain two different distributions of firing rate differences, Δr^{same} and Δr^{diff}, representing the neural discriminability of same or different tokens of textures, and quantify that discriminability with ROC analysis, as above, and calculate

$$\rho_{\text{Token}} = corr[\Delta t, AUC_{\text{Token}}(\Delta t)] \quad (50)$$

in an analogous way to ρ_{Type} as described above.

6.6 Results

6.6.1 Online stimulus optimization

We successfully applied our synthetic texture online evolutionary optimization algorithm to neurons throughout the non-primary auditory cortex. An example evolutionary optimization is shown in Figure 20. Stimuli in initial generations generally evoked low firing rates, even though their center frequencies and intensities were tailored to each neuron. Mutation served to gradually increase firing rates by optimizing stimulus features over the course of a stimulus optimization session. In general, elite textures tended to evoke similar firing rates as their predecessors, suggesting that summary statistics may indeed be a sufficient representation of a sound identity. ANOVA revealed that, of the 45 neurons in our dataset, 41 showed a significant main effect of stimulus origin (elite, mutated or new), and in post-hoc analysis exhibited the highest firing rates to the elite textures, compared to the other two origins. Furthermore, 38/45 neurons showed a significant effect of generation, and had a positive generation coefficient, confirming that the evolutionary algorithm was successful in optimizing firing rates.

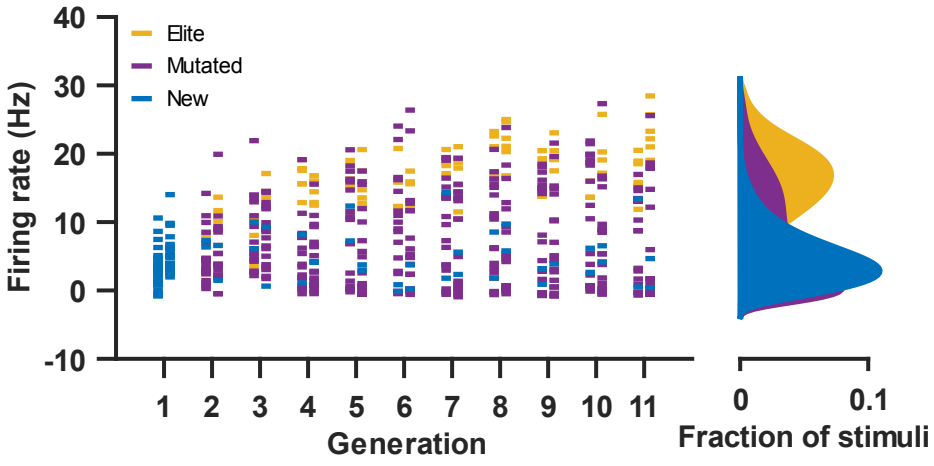


Figure 20 - Example single neuron evolutionary optimization

Left: Each point represents the average firing rate to one synthetic auditory texture stimulus. The two columns of data for each generation indicate the two different lineages. Color indicates the source of each stimuli; blue: generated randomly; violet: mutated from a stimulus in a preceding generation by taking a random step in texture parameter space; yellow: generated by reusing the same texture or parameter values as a stimulus from a preceding generation. **Right:** kernel density estimates of the marginal distributions of firing rates by stimulus category, collapsed over generation and lineage. The highest firing rates are found in the elite (yellow) category.

The single main impediment to applying this technique was the ability to hold neurons for long enough to optimize their stimuli. Running both lineages to ten generations required on the order of two hours. Figure 21 shows a summary of recording time and maximum evoked firing rate over all 87 neurons we attempted this technique on. We selected for further analysis all neurons for which we recorded at least five generations in both lineages, and at least one stimulus in either lineage evoked a firing rate more than 20 Hz above the spontaneous rate. This resulted in a dataset of 45 neurons.

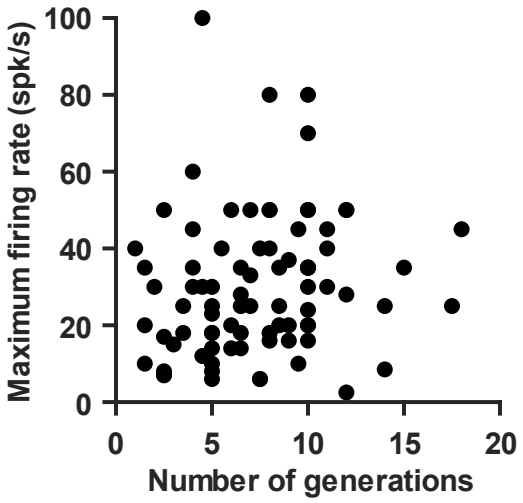


Figure 21 - Summary of all genetic algorithm optimization attempts.

Each point represents one neuron. The x-axis is the maximum number of generations the optimization algorithm ran; the y-axis is the maximum firing evoked (i.e., the average firing rate evoked by the best stimulus).

6.6.2 Modeling the texture receptive field

To summarize how neurons responded to the array of synthetic textures presented by the online stimulus optimization, we constructed for each neuron a 'texture receptive field' (TRF), a linear model that related the texture representation of a sound to neural firing rates. To validate these models, TRFs were fit separately to data from each lineage and used to predict data from the other lineage. Figure 22 shows the model validation results for one example neuron. Models trained on data from the first lineage do well in predicting results from the other.

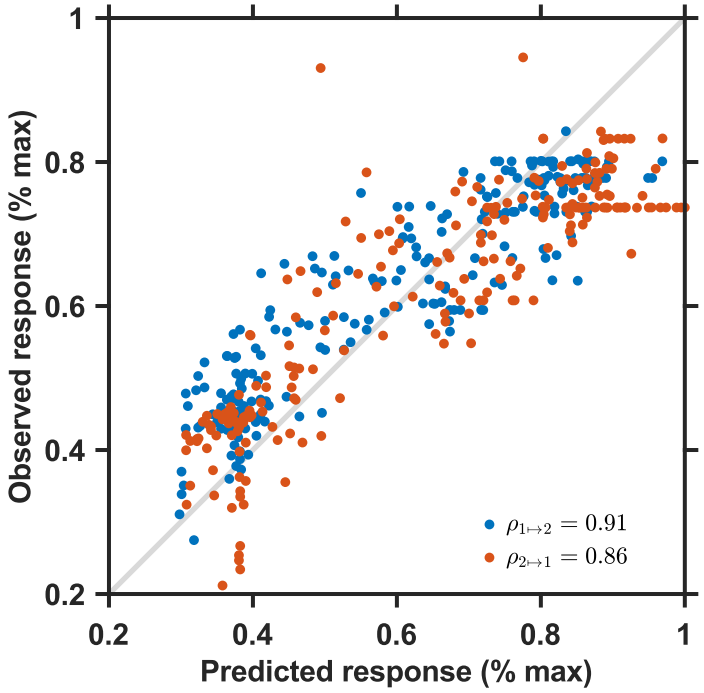


Figure 22 - Texture receptive field regression results for one example neuron.

Each point represents one stimulus, colored by which of the two independent lineages it originated from. Points fall on the diagonal if they are perfectly predicted by the model. The legend indicates the correlation coefficients between predicted and observed responses for both lineages.

Figure 23A shows the distribution of $\bar{\rho}$ for all neurons selected for analysis. 36/45 (80%) of the neurons exhibited significant correlation in both lineage $1 \mapsto 2$ and $2 \mapsto 1$ regression models. In order to provide a comparison to more traditional models of neural receptive fields, we also fit STRF models in a similar way. Figure 23B shows a direct comparison between models in terms of their R^2 values. The TRF models outperform the STRF models (paired t-test, $p < 10^{-4}$.)

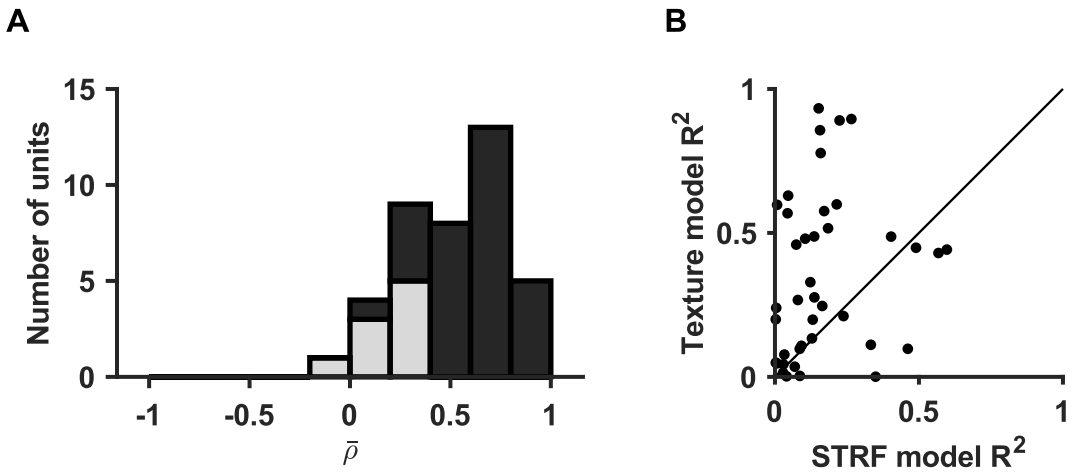


Figure 23 - Distribution of average correlation coefficients $\bar{\rho}$ for fitting neuron texture receptive fields

A: Distribution of $\bar{\rho}$, the correlation between observed responses, and responses predicted by the TRF, averaged over both lineages for each neuron. Dark bars indicate neurons for which the $\rho_{1\mapsto 2}$ and $\rho_{2\mapsto 1}$ model fits were both significant; light bars those where one or both fits were insignificant. **B**: Comparison of R^2 values between the TRF and STRF models. Each point represents one neuron. The TRF model outperforms the STRF model for most neurons. R^2 values are normalized as explained in the methods section.

6.6.3 Classification of real-world textures by a model neural population

What components of the texture model are represented in neural responses? For example, a purely spectral representation might only be affected by the frequency spectrum of the stimuli and ignore any mutations in modulation or envelope correlation. One way to address this question is to manipulate synthetic textures in controlled ways and see how the response changes. If we had an unlimited amount of experimental time with each neuron, we could test this directly, by first finding a synthetic texture that

robustly drove the neuron, and then testing every possible version of that texture with or without each statistical component. One advantage of constructing the texture receptive field model is that we can perform the analogous experiments offline, using the model to predict changes in response to the same texture perturbations.

To help ensure the TRF captured response sensitivity to these sorts of perturbations, the online stimulus optimization process included specially designed perturbed stimuli ('texture controls', see the "Subsequent generations" section of the methods) that removed some of the constraints on the texture specification that did not contribute offspring to subsequent generations.

To quantify these changes, we constructed a virtual neural population and asked how it was able to discriminate real world textures with or without controlled texture perturbations. The virtual neural population consisted of the neurons from the previous section to which we were able to fit a statistically significant texture receptive field. We assumed that neurons responded to a virtually presented texture field independently of each other, with a mean response specified by the TRF, plus an additional independent noise component. We first used these models to generate virtual responses to textures measured from a real-world sound bank and trained a linear discriminator to classify textures based on the population response vector. Next, we predicted neural responses to 'perturbed' textures that were synthesized with or without constraints on different subsets of the statistical structures present in the original sounds and asked

how well the linear discriminator could classify textures missing these statistical components.

The results of the virtual neuron population classifier are shown in Figure 24. In general, the observed trends very closely match the trends seen in the human psychophysics results of McDermott & Simoncelli, 2011. Like humans, the model performs worst, but above chance, when classifying sounds based only on their power spectra, and performs best on textures that include higher order statistical structures like envelope correlations. Compared to the TRF model classification results, the STRF model population performs even worse, and is much less sensitive to the removal of different statistical components of the texture model.

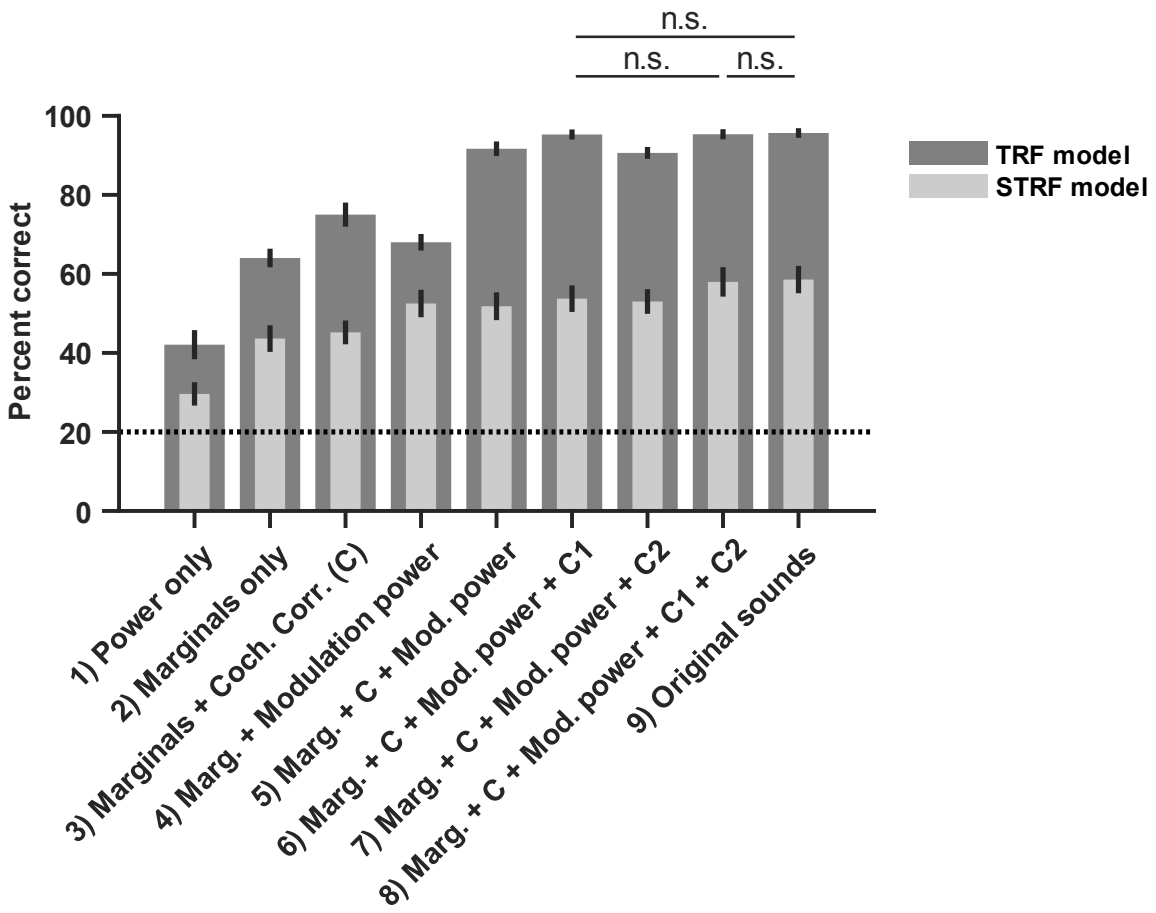


Figure 24 - Virtual neural population texture classification results

Y-Axis indicates the performance of the linear classifier on virtual population response vectors constructed from either the TRF models (dark bars), or STRF models (light bars), trained on textures measured from real world sounds. The dotted line indicates chance level performance. Only non-significant pair-wise comparisons between conditions in the TRF model are indicated; all other TRF pairwise comparisons are significant.

Note that the human psychophysics results in McDermott & Simoncelli, 2011 exhibit a small decrement in performance between real sounds and synthesized textures using the full model, whereas our results do not. (Figure 24, column 8 vs 9) This is presumably because the decrement in human results reflects not only synthesis errors, but also the external validity of the texture model as a whole; that is, the texture model can capture most, but not all aspects of auditory texture perception. Our results measure only internal validity, and any difference between real sounds and sounds synthesized to match real statistics would solely be due to failures to synthesize the correct matching statistics.

6.6.4 Texture discrimination

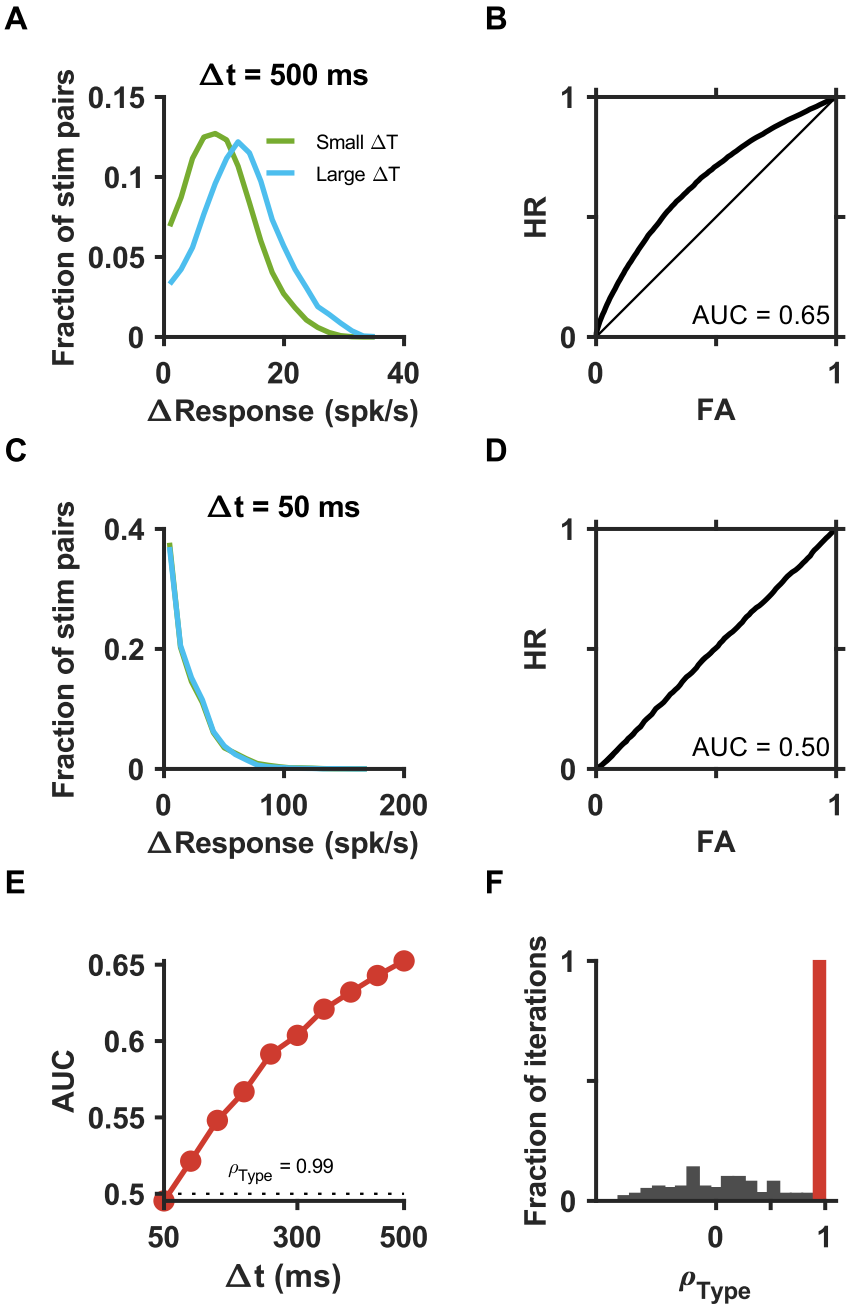
The representation of auditory textures by their summary statistics implies that, by averaging over longer and longer durations, their statistical summaries should approach the true underlying statistical distributions. This was shown to occur in human psychophysics, where, as exemplar duration increased, subjects got progressively better at discriminating synthetic tokens from different texture classes. In contrast, their ability to discriminate different tokens synthesized to match the same texture class progressively *worsened.*

To determine if an analogous effect occurs at the single neuron level we applied an analysis that calculated firing rates across time windows of varying duration. These analyses are shown for an example neuron for 'type' discrimination, the ability to discriminate, based on firing rate differences, whether two different texture stimuli

belong to similar or dissimilar textures, in Figure 25. This neuron's ability to discriminate different types of texture improves with time window duration (Figure 25E, $\rho_{Type} > 0$). In contrast, this neuron's ability to discriminate whether two individual stimulus presentations belong to the same or different token decreases with time window duration (Figure 26E, $\rho_{Token} < 0$). This pattern of positive ρ_{Type} and negative ρ_{Token} is exactly what would be predicted if neurons were representing time-averaged statistical summaries of sound textures, rather than, for example, the specific temporal patterns of each token's time waveform. How common is this pattern at a population level? We summarized our dataset by plotting each neuron as a point specified by its $(\rho_{Type}, \rho_{Token})$ values, shown in Figure 27. Of the 37 neurons in our dataset, 31 of them (84%) fell in the fourth quadrant, meaning they exhibited the same pattern of positive ρ_{Type} and negative ρ_{Token}.

Figure 25 – Example single neuron discrimination between texture classes as a function of stimulus duration.

A: Distributions of response difference Δr between pairs of stimuli that have either similar textures (small ΔT) or dissimilar textures (large ΔT) when using a Δt = 500 ms response window. B: ROC curve for the distributions in A. Because the area under the curve is larger than 0.5, the distributions can be discriminated. C: Same conventions as in A but for Δt = 50 ms time windows. There is no longer any difference between the distributions. D: ROC curve for the distributions in C. E: The relationship between time window duration Δt and the area under the ROC curve. F: Red bars show the distribution of ρ_{Type} obtained by repeating the analysis described by A-D, and randomly choosing the time windows. The grey bars show the distribution of ρ_{Type} values obtained when repeating the same analysis after shuffling the stimulus pairs between the small and large ΔT stimulus pair datasets. The observed and null distributions of ρ_{Type} are significantly different. ($p < 10\text{-}5$)


A
Δt = 500 ms
Fraction of stim pairs
0.15
0.1
0.05
0
Small ΔT
Large ΔT
0
20
40
ΔResponse (spk/s)
B
HR
1
0
AUC = 0.65
0
1
FA
C
Δt = 50 ms
Fraction of stim pairs
0.4
0.2
0
0
100
200
ΔResponse (spk/s)
D
HR
1
0
AUC = 0.50
0
1
FA
E
AUC
0.65
0.6
0.55
0.5
ρType = 0.99
50
300
500
Δt (ms)
F
Fraction of iterations
1
0
0
1
ρType

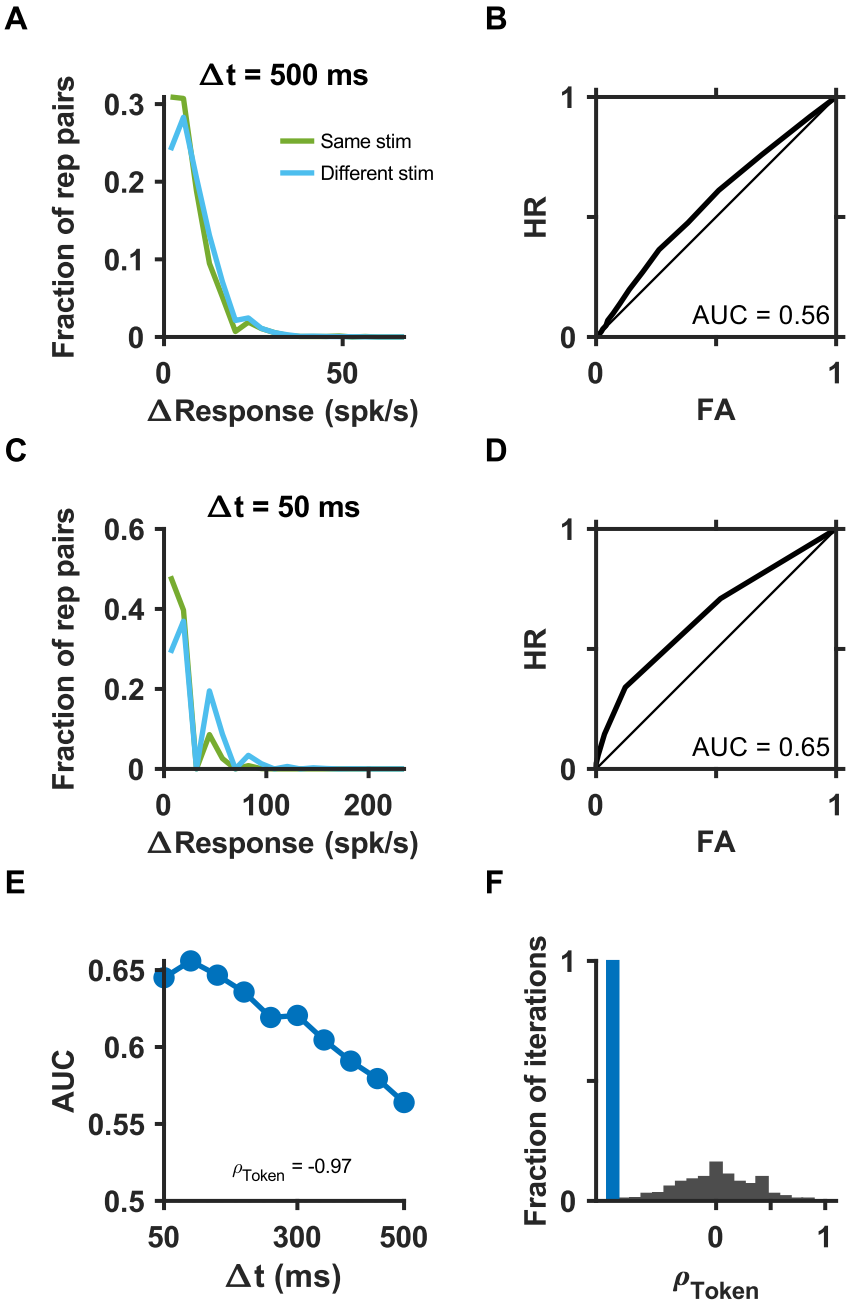


Figure 26 - Example single neuron discrimination between texture tokens as a function of stimulus duration

Same conventions as in Figure 25 but for discrimination of different individual repetitions of either the same token or different tokens of similar textures.

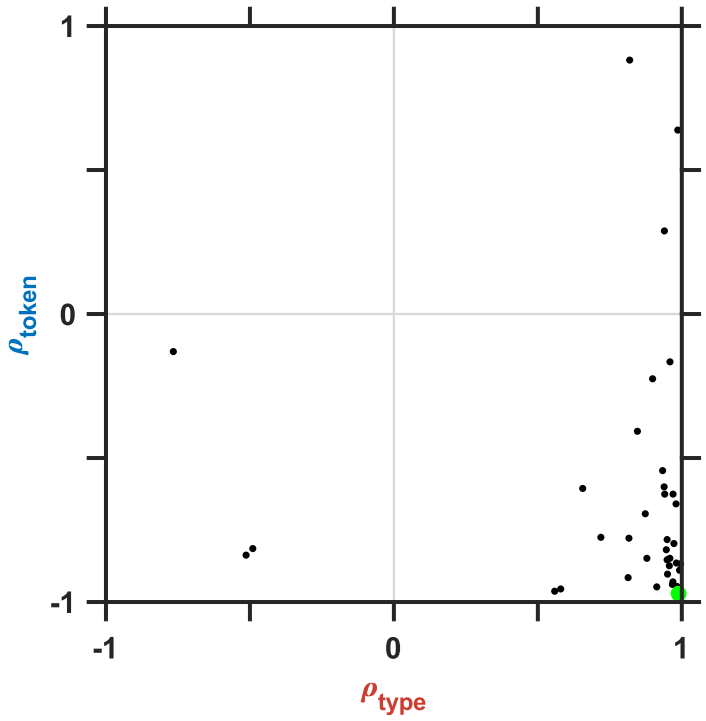


Figure 27 - Summary of type and token discriminability values

Each point represents one neuron by its median ρ_{Type} and ρ_{Token} values. The point highlighted in green represents the neuron illustrated by the example analysis in Figure 25 and Figure 26.

6.7 Conclusion

Auditory textures are a rich class of sound stimuli, with complex spectral and temporal structure, and play an important role in real life auditory stream segregation and auditory navigation. However, the amount of experimental time and the high dimensionality of the stimulus space necessarily limits the ability to thoroughly test responses to these stimuli. We addressed this difficulty in two ways. First, using an online stimulus optimization algorithm modeled after genetic optimization, and a novel synthetic texture parameterization and synthesis algorithm, we were able to more efficiently sample the stimulus space, by concentrating on the regions where the neuron responded. Secondly, we were able to summarize these responses by a novel 'texture receptive field' model.

Fitting a model at the level of the texture specification of a sound was particularly useful because it let us predict responses to textures that were not played during the actual experiments. By first fitting models to data from separate lineages, and validating predictions across lineages, we confirmed the ability of these models to predict responses to independent texture stimuli. By combining single neuron TRFs to form a virtual population, we developed a simple but powerful texture classification approach that was sensitive to different components of the statistical representation in a manner that recapitulated the patterns of classification in human psychophysical experiments.

That neurons in non-primary auditory cortex are sensitive to the power and modulation spectra is not particularly surprising. But particularly interesting is that adding envelope correlation structure, either in terms of the overall C correlations, or the modulation sub-band C1 correlations, improves classifier performance relative to textures that do not contain those structures. (Figure 24; 4 vs. 5, 5 vs 6, 7 vs. 8). These forms of envelope correlations are often termed 'comodulation'. The presence of comodulation in background sound improves target detection in human psychophysical experiments, an effect known as comodulation masking release (CMR). (Hall, Haggard, & Fernandes, 1984; Schooneveldt & Moore, 1989) Neural correlates of CMR have been found in cat auditory cortex. (Nelken, Rotman, & Yosef, 1999) Neurons in marmoset primary auditory cortex have been found sensitive to temporal incoherence. (Barbour & Wang, 2002) Temporal coherence is fundamental to auditory scene analysis. (Shamma, Elhilali, & Micheyl, 2010).

A salient difference between our results and human psychophysics is that we were unable to detect a sensitivity to the C2 correlation structures. Adding C2 statistics to simpler models either had no effect (Figure 24; 6 vs 8), or decreased performance (Figure 24; 5 vs 7). This was surprising given that previous work has shown sensitivity to temporal structure beyond modulation rate (Fishbach, Nelken, & Yeshurun, 2001; Lu et al., 2001; Zhou & Wang, 2010), especially in distinguishing onset and offset transients. (Phillips, Hall, & Boehnke, 2002) This may be due to our texture search strategy, which relies on a highly dimensionally reduced representation of the C2 structure, and likely did not sufficiently search the stimulus space. (See methods).

Further work is required in the analysis of the structure of the C2 statistics to search this space more efficiently in the future.

Why might we expect auditory cortical neurons to represent the statistical structure of sounds? First, as previously described, the statistical measurements used in this model are relatively simple and are easy to measure in a biologically plausible network. More importantly, by representing the statistical structure of sound, the network represents *predictions* about upcoming sounds. This would allow auditory cortex to *subtract* sound textures from its representation of an acoustic scene, which would amplify *unpredicted* sounds; those unpredictable sounds are more likely to represent ethologically relevant and important acoustic signals, such as vocalizations. (Rabinowitz & King, 2011) This line of reasoning suggests several intriguing avenues for future work, especially for target detection in the presence of texture-like background noise.

If sound textures are represented in the brain by their summary statistics, then as stimulus durations increase, their observed summary statistics will approach the true underlying source statistics. This implies that texture class discrimination would improve as stimulus duration increases, and also that texture token discrimination would worsen. This was shown to be the case in human psychophysical experiments (McDermott et al., 2013). We established here that similar trends occur at the single neuron level when using time window averaged firing rates rather than human

classification; a necessary (albeit insufficient) result if auditory cortex is the locus for the representation of texture summary statistics.

There are several caveats in this analysis that should be acknowledged. First, we did not actually vary the duration of the stimuli used in our experiments; we only simulated it by calculating firing rates over time windows shorter than the actual stimulus, randomly positioned within the stimulus presentation window. Secondly, because we used completely arbitrary and random synthetic sound textures, the concept of texture 'class' is not clearly defined, compared to human-labeled sound categories. Instead, we defined a distance function between two different textures and considered pairs of textures that were nearby as belong to the 'same' class. There is some reason to believe the distance function we used is meaningful, insofar as it was successful in synthesizing realistic sound textures when used as the cost function for the gradient descent synthesis algorithm. Further work into human texture perception could investigate a perceptual distance function between synthetic auditory textures.

7. Future Directions

Our experiments found that parabelt regions of marmoset auditory cortex are well-suited to represent the statistical structure of background noise. However, because our behavioral task was designed to simply compare the neural representation between behaving and passive conditions it did not specifically require the discrimination of textures. A major requirement for further studying the neural representation of noise structure in the marmoset auditory cortex would be confirming that marmosets perceive auditory textures in a manner similar to humans. By altering the task to require discriminating along some parameter of the texture space, such as a same/different texture task, it would be possible to compare marmoset and human texture discrimination functions. This would also allow varying stimulus duration to check if marmosets exhibit the same pattern of behavior as in humans where type discrimination improves, and token discrimination worsens with increasing stimulus duration.

Selective attention has been well studied in visual cortex (Desimone & Duncan, 1995; Harris & Thiele, 2011b; Reynolds & Chelazzi, 2004), but much less so in auditory cortex. One of the major difficulties in using selective attention frameworks in the auditory domain is the difficulty in training non-human primates in a selective attention task. One major future direction for experiments in non-primary auditory cortex would be some form of selective attention or stream segregation task, rather than the basic single stream oddball detection task described in this thesis. This would require some way of presenting multiple acoustic streams, separated by space, time, frequency, or

modulation, and some way of cueing the animal to attend to one of the streams. The acoustic textures used in our experiments here are well suited to target-in-noise style tasks, where the goal is to detect oddballs in the target stream, and difficultly is controlled by manipulating the level of the background masking stream. A design like this would also allow the ability to manipulate the statistical structure of the background texture stream in order to control how predictable it is. If our hypotheses about the role of the representation of texture statistics for background subtraction is correct, more reliable statistics (or longer stimulus presentation times allowing more reliable estimation of the sound statistics) should be correlated with improved target detection.

Analyzing neural firing rates and relating them to behavioral task performance can only provide correlations. It would be a major step forward to assign a causal role to any of the non-primary auditory fields by using some form of manipulation to observe disruptions in behavioral performance. In particular, the hypothesized dissociation between spatial and non-spatial tasks in the dual stream model has only ever been causally manipulated in the cat (Lomber & Malhotra, 2008), where the cortical cooling of either the anterior or posterior auditory fields AAF and PAF resulted in doubly dissociated performance decrements in either temporal discrimination task, or a spatial task. This was interpreted as support for the dual stream hypothesis. However, the fact that the tonotopy of A1 is flipped relative to the primate A1 makes it difficult to say whether PAF and AAF are homologous to the caudal and rostral auditory fields in primate auditory cortex.

If the high frequency tonotopic reversal in parabelt auditory cortex we have identified is correct, it would provide a convenient landmark for future interrogations of parabelt function, where the two regions are independently manipulated in a dual dissociation design. One appropriate design would be a target detection in noise. The noise would be constructed in a similar manner to those used in the current experiments, where their statistical structure is controlled. Targets would be simple foreground sounds that change in some acoustic parameters such as location or pitch. A cue would indicate which acoustic parameter to attend to changes in. Our prediction is that selective interference in the two different parabelt fields will induce selective behavioral changes depending on which dimensions of attention the task is demanding, so for example in trials where attention is cued to spatially allocated targets, inactivation of causal parabelt would result in behavior deficits, but not rostral parabelt.

Bibliography

Atiani, S., David, S. V, Elgueda, D., Locastro, M., Radtke-Schuller, S., Shamma, S. a, & Fritz, J. B. (2014). Emergent selectivity for task-relevant stimuli in higher-order auditory cortex. *Neuron*, *82*(2), 486–99. http://doi.org/10.1016/j.neuron.2014.02.029

Barbour, D. L., & Wang, X. (2002). Temporal coherence sensitivity in auditory cortex. *Journal of Neurophysiology*, *88*(5), 2684–99. http://doi.org/10.1152/jn.00253.2002

Barbour, D. L., & Wang, X. (2003). Auditory cortical responses elicited in awake primates by random spectrum stimuli. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *23*(18), 7194–206. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1945239&tool=pmcentrez&rendertype=abstract

Baumann, S., Joly, O., Rees, A., Petkov, C. I., Sun, L., Thiele, A., & Griffiths, T. D. (2015). The topography of frequency and time representation in primate auditory cortices. *ELife*, *4*, 1–15. http://doi.org/10.7554/eLife.03256

Baumann, S., Petkov, C. I., & Griffiths, T. D. (2013). A unified framework for the organization of the primate auditory cortex. *Frontiers in Systems Neuroscience*, *7*(11). http://doi.org/10.3389/fnsys.2013.00011

Bendor, D., & Wang, X. (2008). Neural response properties of primary, rostral, and rostrotemporal core fields in the auditory cortex of marmoset monkeys. *Journal of Neurophysiology*, *100*(2), 888–906. http://doi.org/10.1152/jn.00884.2007

Bregman, A. S. (1990). *Auditory Scene Analysis*. Cambridge, MA: MIT Press.

Brosch, M., Selezneva, E., & Scheich, H. (2005). Nonauditory Events of a Behavioral Procedure Activate Auditory Cortex of Highly Trained Monkeys. *J. Neurosci.*, *25*(29), 6797–6806. http://doi.org/10.1523/JNEUROSCI.1571-05.2005

Brosch, M., Selezneva, E., & Scheich, H. (2011). Representation of reward feedback in primate auditory cortex. *Frontiers in Systems Neuroscience*, *5*(February), 5. http://doi.org/10.3389/fnsys.2011.00005

Camalier, C., D'Angelo, W., Sterbing-D'Angelo, S., Mothe, L. de la, & Hackett, T. (2012). Neural latencies across auditory cortex of macaque support a dorsal stream supramodal timing advantage in primates. *PNAS*, *109*(44), 18168–73. http://doi.org/10.1073/pnas.1206387109

Carlson, E. T., Rasquinha, R. J., Zhang, K., & Connor, C. E. (2011). A sparse object coding scheme in area V4. *Current Biology*, *21*(4), 288–93. http://doi.org/10.1016/j.cub.2011.01.013

Chambers, A. R., Hancock, K. E., Sen, K., & Polley, D. B. (2014). Online stimulus optimization rapidly reveals multidimensional selectivity in auditory cortical neurons. *The Journal of Neuroscience*, *34*(27), 8963–75. http://doi.org/10.1523/JNEUROSCI.0260-14.2014

Chase, S. M., & Young, E. D. (2007). First-spike latency information in single neurons increases when referenced to population onset. *Proceedings of the National Academy of Sciences*, *104*(12), 5175–80. http://doi.org/10.1073/pnas.0610368104

Crochet, S., & Petersen, C. C. H. (2006). Correlating whisker behavior with membrane potential in barrel cortex of awake mice, *9*(5), 608–610.

http://doi.org/10.1038/nn1690

de la Mothe, L. a, Blumell, S., Kajikawa, Y., & Hackett, T. a. (2012). Cortical connections of auditory cortex in marmoset monkeys: lateral belt and parabelt regions. *Anatomical Record*, *295*(5), 800–21. http://doi.org/10.1002/ar.22451

Depireux, D. a, Simon, J. Z., Klein, D. J., & Shamma, S. a. (2001). Spectro-temporal response field characterization with dynamic ripples in ferret primary auditory cortex. *Journal of Neurophysiology*, *85*(3), 1220–1234.

Desimone, R., & Duncan, J. (1995). Neural Mechanisms of Selective Visual Attention. *Annual Review of Neuroscience*, *18*, 193–222.

Fishbach, A., Nelken, I., & Yeshurun, Y. (2001). Auditory edge detection: a neural model for physiological and psychoacoustical responses to amplitude transients. *Journal of Neurophysiology*, *85*(6), 2303–23. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11387378

Fritz, J. B., David, S. V, Radtke-Schuller, S., Yin, P., & Shamma, S. A. (2010). Adaptive, behaviorally gated, persistent encoding of task-relevant auditory information in ferret frontal cortex. *Nat Neurosci*, *advance on*. http://doi.org/10.1038/nn.2598

Fritz, J. B., Elhilali, M., & Shamma, S. A. (2005). Differential dynamic plasticity of A1 receptive fields during multiple spectral tasks. *Journal of Neuroscience*, *25*(33), 7623.

Fritz, J. B., Elhilali, M., & Shamma, S. A. (2007). Adaptive Changes in Cortical Receptive Fields Induced by Attention to Complex Sounds. *J Neurophysiol*, *98*(4), 2337–2346. http://doi.org/10.1152/jn.00552.2007

Fritz, J., Shamma, S., Elhilali, M., & Klein, D. (2003). Rapid task-related plasticity of spectrotemporal receptive fields in primary auditory cortex. *Nat Neurosci*, *6*(11), 1216–1223. http://doi.org/10.1038/nn1141

Goldberg, J. M., & Brown, P. B. (1969). Response of binaural neurons of dog superior olivary complex to dichotic tonal stimuli: some physiological mechanisms of sound localization. *Journal of Neurophysiology*, *32*(4), 613–636. http://doi.org/10.1007/978-1-4612-2700-7_3

Goodale, M., & Milner, D. (1992). Separate visual pathways for perception and action. *Trends in Neurosciences1*, *15*(1), 20–25.

Hackett, T., Stepniewska, I., & Kaas, J. H. (1998). Subdivisions of auditory cortex and ipsilateral cortical connections of the parabelt auditory cortex in macaque monkeys. *The Journal of Comparative Neurology*, *394*(4), 475–95. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9590556

Hall, J. W., Haggard, M. P., & Fernandes, M. A. (1984). Detection in noise by spectro-temporal pattern analysis. *The Journal of the Acoustical Society of America*, *76*(1), 50–56. http://doi.org/10.1121/1.391005

Harris, K. D., & Thiele, A. (2011a). Cortical state and attention. *Nature Reviews. Neuroscience*, *12*(September). http://doi.org/10.1038/nrn3084

Harris, K. D., & Thiele, A. (2011b). Cortical state and attention. *Nature Reviews. Neuroscience*, *12*(9), 509–23. http://doi.org/10.1038/nrn3084

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning* (2nd ed.). Springer.

Hung, C. C., Carlson, E. T., & Connor, C. E. (2012). Medial Axis Shape Coding in

Macaque Inferotemporal Cortex. *Neuron*, *74*(6), 1099–1113. http://doi.org/10.1016/j.neuron.2012.04.029

Kaas, J. H., & Hackett, T. A. (2000). Subdivisions of auditory cortex and processing streams in primates. *Proceedings of the National Academy of Sciences of the United States of America*, *97*(22), 11793–11799. http://doi.org/VL - 97

Kaas, J. H., & Hackett, T. a. (1999). "What" and "where" processing in auditory cortex. *Nature Neuroscience*, *2*(12), 1045–7. http://doi.org/10.1038/15967

Kajikawa, Y., Frey, S., Ross, D., Falchier, A., Hackett, T. a, & Schroeder, C. E. (2015). Auditory properties in the parabelt regions of the superior temporal gyrus in the awake macaque monkey: an initial survey. *Journal of Neuroscience*, *35*(10), 4140–50. http://doi.org/10.1523/JNEUROSCI.3556-14.2015

Klein, D. J., Depireux, D. A., Simon, J. Z., & Shamma, S. A. (2000). Robust Spectrotemporal Reverse Correlation for the Auditory System: Optimizing Stimulus Design. *Journal of Computational Neuroscience*, *9*(1), 85–111. http://doi.org/10.1023/A:1008990412183

Lee, C.-C., & Middlebrooks, J. C. (2010). Auditory cortex spatial sensitivity sharpens during task performance. *Nature Neuroscience*, *14*(1), 108–114. http://doi.org/10.1038/nn.2713

Liang, L., Lu, T., & Wang, X. (2002). Neural representations of sinusoidal amplitude and frequency modulations in the primary auditory cortex of awake primates. *Journal of Neurophysiology*, *87*(5), 2237–61. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11976364

Lomber, S. G., & Malhotra, S. (2008). Double dissociation of "what" and "where"

processing in auditory cortex. *Nature Neuroscience*, *11*(5), 609–16. http://doi.org/10.1038/nn.2108

Lu, T., Liang, L., & Wang, X. (2001). Temporal and rate representations of time-varying signals in the auditory cortex of awake primates. *Nature Neuroscience*, *4*(11), 1131–8. http://doi.org/10.1038/nn737

Mardia, K., & Jupp, P. (2000). *Directional Statistics*. New York: Wiley.

McDermott, J. H., Schemitsch, M., & Simoncelli, E. P. (2013). Summary statistics in auditory perception. *Nature Neuroscience*, *16*(4), 493–498. http://doi.org/10.1038/nn.3347

McDermott, J. H., & Simoncelli, E. P. (2011). Sound texture perception via statistics of the auditory periphery: Evidence from sound synthesis. *Neuron*, *71*(5), 926–940. http://doi.org/10.1016/j.neuron.2011.06.032

McGinley, M. J., Vinck, M., Reimer, J., Batista-Brito, R., Zagha, E., Cadwell, C. R., … McCormick, D. A. (2015). Waking State: Rapid Variations Modulate Neural and Behavioral Responses. *Neuron*, *87*(6), 1143–1161. http://doi.org/10.1016/j.neuron.2015.09.012

Mishkin, M., Ungerleider, L. G., & Macko, K. A. (1983). Object vision and spatial vision: two cortical pathways. *Trends in Neurosciences*, *6*, 414–417. http://doi.org/10.1016/0166-2236(83)90190-X

Moerel, M., De Martino, F., & Formisano, E. (2014). An anatomical and functional topography of human auditory cortical areas. *Frontiers in Neuroscience*, *8*(225). http://doi.org/10.3389/fnins.2014.00225

Moran, J., & Desimone, R. (1985). Selective attention gates visual processing in the

extrastriate cortex. *Science*, *229*. Retrieved from http://www.sciencemag.org/content/229/4715/782.short

Nelken, I., Rotman, Y., & Yosef, O. B. (1999). Responses of auditory-cortex neurons to structural features of natural sounds. *Nature*, *397*(6715), 154–157. http://doi.org/10.1038/16456

Niell, C. M., & Stryker, M. P. (2010). Modulation of Visual Responses by Behavioral State in Mouse Visual Cortex. *Neuron*, *65*(4), 472–479. http://doi.org/10.1016/j.neuron.2010.01.033

O'Connor, D. H., Peron, S. P., Huber, D., & Svoboda, K. (2010). Neural Activity in Barrel Cortex Underlying Vibrissa-Based Object Localization in Mice. *Neuron*, *67*(6), 1048–1061. http://doi.org/10.1016/j.neuron.2010.08.026

Okazawa, G., Tajima, S., & Komatsu, H. (2015). Image statistics underlying natural texture selectivity of neurons in macaque V4. *Proceedings of the National Academy of Sciences*, *112*(4), E351–E360. http://doi.org/10.1073/pnas.1415146112

Ortiz-Rios, M., Azevedo, F. A. C., Kuśmierek, P., Balla, D. Z., Munk, M. H., Keliris, G. A., … Rauschecker, J. P. (2017). Widespread and Opponent fMRI Signals Represent Sound Location in Macaque Auditory Cortex. *Neuron*, 971–983. http://doi.org/10.1016/j.neuron.2017.01.013

Osmanski, M. S., & Wang, X. (2011). Measurement of absolute auditory thresholds in the common marmoset (Callithrix jacchus). *Hearing Research*, *277*(1–2), 127–33. http://doi.org/10.1016/j.heares.2011.02.001

Perrodin, C., Kayser, C., Logothetis, N. K., & Petkov, C. I. (2011). Voice cells in the

primate temporal lobe. *Current Biology*, *21*(16), 1408–1415. http://doi.org/10.1016/j.cub.2011.07.028

Petkov, C. I., Kayser, C., Steudel, T., Whittingstall, K., Augath, M., & Logothetis, N. K. (2008). A voice region in the monkey brain. *Nature Neuroscience*, *11*(3), 367–74. http://doi.org/10.1038/nn2043

Phillips, D. P., Hall, S. E., & Boehnke, S. E. (2002). Central auditory onset responses, and temporal asymmetries in auditory perception. *Hearing Research*, *167*(1–2), 192–205. http://doi.org/10.1016/S0378-5955(02)00393-3

Poirier, C., Baumann, S., Dheerendra, P., Joly, O., Hunter, D., Balezeau, F., … Griffiths, T. D. (2017). Auditory motion-specific mechanisms in the primate brain. *PLoS Biology*, *15*(5), 1–24. http://doi.org/10.1371/journal.pbio.2001379

Polack, P., Friedman, J., & Golshani, P. (2013). Cellular mechanisms of brain state – dependent gain modulation in visual cortex. *Nature Neuroscience*, (July), 1–11. http://doi.org/10.1038/nn.3464

Poremba, A., Malloy, M., Saunders, R. C., Carson, R. E., Herscovitch, P., & Mishkin, M. (2004). Species-specific calls evoke asymmetric activity in the monkey's temporal poles. *Nature*, *427*(6973), 448–451. http://doi.org/10.1038/nature02268

Rabinowitz, N. C., & King, A. J. (2011). Auditory perception: Hearing the texture of sounds. *Current Biology*, *21*(23), R967–R968. http://doi.org/10.1016/j.cub.2011.10.027

Rauschecker, J. P., & Tian, B. (2004). Processing of band-passed noise in the lateral auditory belt cortex of the rhesus monkey. *Journal of Neurophysiology*, *91*(6), 2578–89. http://doi.org/10.1152/jn.00834.2003

Rauschecker, J. P., Tian, B., Pons, T., & Mishkin, M. (1997). Serial and parallel processing in rhesus monkey auditory cortex. *The Journal of Comparative Neurology*, *382*(1), 89–103.

Rauschecker, J., Tian, B., & Hauser, M. (1995). Processing of complex sounds in the macaque nonprimary auditory cortex. *Science*, *268*(5207), 111–114. Retrieved from http://www.sciencemag.org/content/268/5207/111.short

Recanzone, G. H. (2000). Response profiles of auditory cortical neurons to tones and noise in behaving macaque monkeys. *Hearing Research*, *150*(1–2), 104–18. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11077196

Recanzone, G. H., Engle, J. R., & Juarez-Salinas, D. L. (2011). Spatial and temporal processing of single auditory cortical neurons and populations of neurons in the macaque monkey. *Hearing Research*, *271*(1–2), 115–22. http://doi.org/10.1016/j.heares.2010.03.084

Recanzone, G. H., Guard, D. C., Phan, M. L., & Su, T. K. (2000). Correlation between the activity of single auditory cortical neurons and sound-localization behavior in the macaque monkey. *Journal of Neurophysiology*, *83*(Cm), 2723–2739.

Remington, E. D., Osmanski, M. S., & Wang, X. (2012). An operant conditioning method for studying auditory behaviors in marmoset monkeys. *PloS One*, *7*(10), e47895. http://doi.org/10.1371/journal.pone.0047895

Remington, E. D., & Wang, X. (2019). Neural Representations of the Full Spatial Field in Auditory Cortex of Awake Marmoset (Callithrix jacchus). *Cerebral Cortex*, *29*(March), 1199–1216. http://doi.org/10.1093/cercor/bhy025

Reynolds, J. H., & Chelazzi, L. (2004). Attentional modulation of visual processing.

Annual Review of Neuroscience, *27*, 611–47. http://doi.org/10.1146/annurev.neuro.26.041002.131039

Romanski, L. M., Bates, J. F., & Goldman-Rakic, P. (1999). Auditory belt and parabelt projections to the prefrontal cortex in the rhesus monkey. *J Comp Neurology*, (403), 141–157.

Romanski, L. M., Tian, B., Fritz, J., Mishkin, M., Goldman-Rakic, P. S., & Rauschecker, J. P. (1999). Dual streams of auditory afferents target multiple domains in the primate prefrontal cortex. *Nature Neuroscience*, *2*(12), 1131–6. http://doi.org/10.1038/16056

Sadagopan, S., Temiz-Karayol, N. Z., & Voss, H. U. (2015). High-field functional magnetic resonance imaging of vocalization processing in marmosets. *Scientific Reports*, *5*, 10950. http://doi.org/10.1038/srep10950

Schmolesky, M. T., Wang, Y., Hanes, D., Thompson, K. G., Leutgeb, S., Schall, J. D., & Leventhal, a G. (1998). Signal timing across the macaque visual system. *J Neurophysiol*, *79*(6), 3272–3278. http://doi.org/10.1016/j.actpsy.2013.06.009

Schooneveldt, G. P., & Moore, B. C. J. (1989). Comodulation masking release (CMR) as a function of masker bandwidth, modulator bandwidth, and signal duration. *The Journal of the Acoustical Society of America*, *85*(1), 273–281. http://doi.org/10.1121/1.397734

Shamma, S. a., Elhilali, M., & Micheyl, C. (2010). Temporal coherence and attention in auditory scene analysis. *Trends in Neurosciences*, *34*(3), 114–123. http://doi.org/10.1016/j.tins.2010.11.002

Striem-Amit, E., Hertz, U., & Amedi, A. (2011). Extensive cochleotopic mapping of

human auditory cortical fields obtained with phase-encoding FMRI. *PloS One*, *6*(3), e17832. http://doi.org/10.1371/journal.pone.0017832

Tani, T., Abe, H., Hayami, T., Banno, T., & Miyakawa, N. (2018). Sound Frequency Representation in the Auditory Cortex of the Common Marmoset Visualized Using Optical Intrinsic Signal Imaging, *5*(April), 1–13.

Tian, B., & Rauschecker, J. P. (2004). Processing of frequency-modulated sounds in the lateral auditory belt cortex of the rhesus monkey. *Journal of Neurophysiology*, *92*(5), 2993–3013. http://doi.org/10.1152/jn.00472.2003

Tian, B., Reser, D., Durham, A., Kustov, A., & Rauschecker, J. P. (2001). Functional Specialization in Rhesus Monkey Auditory Cortex. *Science*, *292*(5515), 290–293.

Vaziri, S., Carlson, E. T., Wang, Z., & Connor, C. E. (2014). A channel for 3D environmental shape in anterior inferotemporal cortex. *Neuron*, *84*(1), 55–62. http://doi.org/10.1016/j.neuron.2014.08.043

Vaziri, S., & Connor, C. E. (2016). Representation of gravity-aligned scene structure in ventral pathway visual cortex. *Current Biology*, *26*(6), 766–774. http://doi.org/10.1016/j.cub.2016.01.022

Wang, X., Lu, T., Snider, R. K., & Liang, L. (2005). Sustained firing in auditory cortex evoked by preferred stimuli. *Nature*, *435*(7040), 341–346. http://doi.org/10.1038/nature03565

Woods, T. M., Lopez, S. E., Long, J. H., Rahman, J. E., & Recanzone, G. H. (2006). Effects of stimulus azimuth and intensity on the single-neuron activity in the auditory cortex of the alert macaque monkey. *Journal of Neurophysiology*, *96*(6), 3323–3337. http://doi.org/10.1152/jn.00392.2006

Yamane, Y., Carlson, E. T., Bowman, K. C., Wang, Z., & Connor, C. E. (2008). A neural code for three-dimensional object shape in macaque inferotemporal cortex. *Nature Neuroscience*, *11*(11), 1352–1360. http://doi.org/10.1038/nn.2202

Yin, P., Fritz, J. B., & Shamma, S. a. (2014). Rapid Spectrotemporal Plasticity in Primary Auditory Cortex during Behavior. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *34*(12), 4396–408. http://doi.org/10.1523/JNEUROSCI.2799-13.2014

Yu, J. J., & Young, E. D. (2000). Linear and nonlinear pathways of spectral information transmission in the cochlear nucleus. *Proc Natl Acad Sci U S A*, *97*(22), 11780–11786. http://doi.org/10.1073/pnas.97.22.11780

Zagha, E., & Mccormick, D. A. (2014). Neural control of brain state. *Current Opinion in Neurobiology*, *29*, 178–186. http://doi.org/10.1016/j.conb.2014.09.010

Zhou, Y., & Wang, X. (2010). Cortical Processing of Dynamic Sound Envelope Transitions. *Journal of Neuroscience*, *30*(49), 16741–16754. http://doi.org/10.1523/JNEUROSCI.2016-10.2010

Zhou, Y., & Wang, X. (2012). Level dependence of spatial processing in the primate auditory cortex. *Journal of Neurophysiology*, *108*(3), 810–826. http://doi.org/10.1152/jn.00500.2011

Curriculum Vitae

Darik Gamble

darik.gamble@gmail.com
208 E Biddle St., Apt. 2, Baltimore MD 21202

Education

Johns Hopkins University — Baltimore, MD, USA
- PhD in Biomedical Engineering — Spring 2020

University of Waterloo — Waterloo, ON, Canada
- Bachelor of Applied Science in Chemical Engineerin — 2008
    - Co-operative Program, With Distinction, Dean's Honours List

Research Experience

PhD Student — August 2009 – Present
Dr. Xiaoqin Wang
Johns Hopkins University, Baltimore, MD
- Studying the topographic and functional organization of non-primary auditory cortex with extracellular neurophysiology in head-fixed awake behaving marmoset monkeys

Research Assistant — January 2007 – July 2009
Dr. Eric Jervis
University of Waterloo, Ontario
- Performed long-term *in vitro* live-cell imaging experiments on hematopoeitic stem cells
- Designed and implemented a fully-featured software suite written in Matlab and MySQL for interacting with large (>1 TB) multi-dimensional datasets of live-cell time-lapse video-microscopy.

Research Assistant — September 2005 – April 2006
Dr. Henry Peng
Defence Research & Development Canada, Toronto, Ontario
- Investigated novel strategies for hemorrhage control using thromboelastography.

Manuscripts in Preparation

Gamble D & X Wang. **Physiological characterization of parabelt auditory cortex in the awake, behaving marmoset monkey.**

Gamble D, Kostlan K, Zhang K & X Wang. **Evolutionary optimization of synthetic auditory texture stimuli reveals auditory cortex representation of acoustic statistical structure.**

Refereed Publications

Huang J, Gamble D, Sarnlertsophon K, Wang X, & S Hsiao. 2012. **Feeling music: integration of auditory and tactile inputs in musical meter perception.** PloS One, 7(10)

Moogk, D, Stewart M, Gamble D, Bhatia M & E Jervis. 2010. **Human ESC colony formation is dependent on interplay between self-renewing hESCs and unique precursors responsible for niche generation.** Cytometry A. 77(4), 321–7.

Kachouie N, Fieguth P, Gamble D, Jervis E, Ezziane Z, Khademhosseini A**. Constrained watershed method to infer morphology of mammalian cells in microscopic images.** Cytometry A. 77(12), 1148-1159.

Peng H, Gamble D, and P Shek, **"Thrombelastography (TEG) analysis of hemostatic agents"**, Unclassified Technical Report, TR 2006-252, Defence R & D Canada. (2006)

Conference abstracts

D Gamble and X Wang. **The Representation of Spatial Location and Temporal Modulation in Marmoset Parabelt Auditory Cortex.** Poster presentation, International Conference on Auditory Cortex 2017.

D Gamble and X Wang. **A high frequency tonotopic reversal in marmoset parabelt auditory cortex.** Poster presentation, Society for Neuroscience 2016.

D Gamble and X Wang. **Single-unit characterization of lateral belt and parabelt auditory cortex of the behaving marmoset monkey.** Society for Neuroscience 2015.

D Gamble and X Wang. **Single-unit characterization of putative auditory parabelt cortex in the marmoset monkey.** ARO 2014.

Teaching experience

Teaching assistant - BME 580: Systems bioengineering II

- Undergraduate level neuroscience course

Honors

- ICAC NIH travel award (2017)
- Sandford Fleming Foundation Co-op Proficiency Award (2008)
- Canadian Stem Cell Network Co-op Training Award (2007) $10,000
- Natural Sciences and Engineering Research Council Undergraduate Student Research Award (2007) $2,500
- Faculty of Engineering Upper-Year Scholarship for Academic Achievement (2007) $400
- S.C. Johnson & Son Ltd. Award for Excellence in Written Communication (2006)
  University of Waterloo Undergraduate Research Assistantship (2005) $600